Title: Left-hemisphere cortical language regions respond equally to dialogue and monologue

Abbreviated Title: Language regions respond equally to dialogue and monologue

Authors: Halie Olson[1], Emily Chen[1], Kirsten Lydic[1], Rebecca Saxe[1]

[1] Massachusetts Institute of Technology

Correspondence should be addressed to: Halie Olson (holson@mit.edu)

Acknowledgements: We would like to thank Somaia Saba, Hana Ro, and Michelle Hung for their assistance with stimuli creation. Thank you to Ev Fedorenko, Shari Liu, and Nancy Kanwisher for helpful feedback on this manuscript. Thank you to Steve Shannon and Atsushi Takahashi at the Athinoula A. Martinos Imaging Center at MIT. Finally, thank you to our participants for making this research possible.

Authors report no conflict of interest.

Funding Sources: This research was supported by the Simons Foundation Autism Research Initiative via the Simons Center for the Social Brain at MIT and the NSF Graduate Research Fellowship Program (#1745302 to HO).

Abstract

Much of the language we encounter in our everyday lives comes in the form of conversation, yet the majority of research on the neural basis of language comprehension has used language input from a single source. To determine whether canonical left-hemisphere language regions are sensitive to features of dialogue beyond the comprehensibility of the speech stream, we scanned 20 adults on two novel tasks using functional magnetic resonance imaging. In the first, participants watched videos of puppets speaking either to the viewer (monologue) or to a partner (dialogue), while the audio was either comprehensible (forward) or reversed (backward). Canonical left-hemisphere language regions responded more to forward than backward speech, as expected, but did not respond more to dialogue than monologue. In a second task, two puppets conversed with each other, but only one was comprehensible while the other's speech stream was reversed. Left-hemisphere cortical language regions again responded more to forward than backwards speech, and activity in these regions was only correlated among participants who heard the same characters speaking forward and backward. In contrast, some theory of mind regions and right hemisphere homologues of language regions responded more to dialogue than monologue, and activity in some of these regions was correlated among participants even when opposite characters were speaking forward and backward (in both cases, the visual video clips were held constant). Together, these experiments suggest that canonical left-hemisphere cortical language regions are only sensitive to the language input in dialogue.

Keywords: fMRI, conversation, language comprehension, dialogue, theory of mind

Introduction

Language is first heard, learned and used in informal conversation. By contrast, most research on the neural basis of language comprehension has relied on language from a single source. Compared to monologues or single-source texts, language in turn-taking dialogue exhibits distinctive features that function to coordinate and monitor the creation of common ground (Clark, 1996; Clark & Schaefer, 1989; Fox Tree, 1999; Fusaroli & Tylén, 2016). Successive utterances not only convey new meaning, but often show how a prior utterance was understood, facilitating rapid correction (Schegloff et al., 1977). In conversation, speakers quickly volley back and forth, establishing referents across speaker boundaries and often finishing each other's sentences (Clark, 1996; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986). Speakers alternate about every 2 seconds, with only a 200 ms delay between their utterances on average (Levinson, 2016; Stivers et al., 2009). When observing conversation, adults and even young children can accurately predict turn taking (Casillas & Frank, 2017). Consequently, although utterances in dialogue are typically not well-formed grammatical sentences, dialogue is easier to comprehend than monologue from a single speaker (Fox Tree, 1999).

Here, for example, is a short transcribed excerpt from a two speaker dialogue:

> *Well, you see, I've never met him, and so if he comes to the door, how will I know that it's him? Ah. Oh well, it's easy. For one thing, we're exactly alike. You are? Yeah! We're twins! (Source: https://youtu.be/sS7_-h882Ls)*

As a single linguistic stream, this excerpt is hard to understand, including sentence fragments and seeming disfluencies. Yet when the utterances are assigned to different speakers, the dialogue becomes easily comprehensible:

> ***Ernie: Well, you see, I've never met him, and so if he comes to the door, how will I know that it's him?***
>
> *Bert: Ah. Oh well, it's easy. For one thing, we're exactly alike.*
>
> ***Ernie: You are?***
>
> *Bert: Yeah! We're twins!*

In this example, the backchannel utterance "*Ah*" conveys understanding, and the clarifying question *"You are?"* marks distinct perspectives. Rather than seeming disfluent, these utterances aid in language comprehension (Clark, 1996; Clark & Schaefer, 1989; Tolins & Fox Tree, 2016). Bert knows something about the absent referent that Ernie does not (i.e., that this person is Bert's twin). Representing these different perspectives is integral to understanding the dialogue and predicting what might come next. Comprehending language in dialogue thus requires additional social and linguistic processing, compared to comprehending language from a single source.

Because of this feature, observed dialogue provides an interesting test case for probing the functions of cortical regions involved in language processing. A consistent set of left hemisphere frontal and temporal regions are involved in processing language (Bates et al., 2001; Binder et al., 1997; Dronkers et al., 2004; Fedorenko et al., 2010, 2011; Friederici, 2011; Friederici &

Gierhan, 2013; Price, 2010, 2012). These regions, which we will refer to as the 'canonical language network', robustly respond to language whether it is spoken (Scott et al., 2017), written (Fedorenko et al., 2010), or signed (MacSweeney et al., 2008; Neville et al., 1998). They are active during both production and comprehension (Hagoort, 2014; Hu et al., 2022; Menenti et al., 2011; Price, 2010), in adults and children (Enge et al., 2020), across a wide range of languages (Malik-Moraleda et al., 2022). They are also sensitive to features of language like comprehension difficulty (Wehbe et al., 2021) and syntactic complexity (Blank et al., 2016), responding more to higher syntactic and semantic processing demands (Hagoort & Indefrey, 2014).

Since early lesions studies, it has generally been agreed that canonical language network regions are *necessary* for language (Broca, 1865; Wernicke, 1874). However, there have been long standing debates about the *specificity* of these regions for language processing, and what their limits and scope may be (Fedorenko & Thompson-Schill, 2014; Monti et al., 2012). Initially, whole brain activation mapping suggested that language activates regions that overlapped with a range of other cognitive tasks (Blumstein & Amso, 2013; Gold & Buckner, 2002; Thompson-Schill et al., 1997). When language regions are functionally localized within individuals (Braga et al., 2020; Fedorenko et al., 2010), these regions are not engaged by other compositional or cognitively difficult tasks, like working memory, math, music, cognitive control, action observation, or imitation (Fedorenko et al., 2011; Pritchett et al., 2018). But what about a task that is more cognitively similar? Recent studies looked at a particularly interesting boundary case: reading computer code. Coding shares a number of features with language processing: both, for instance, involve recursively combining components in constrained ways to form a

more complex meaning (Fedorenko et al., 2019). Despite the underlying similarities, language regions are not recruited when people read and evaluate the meaning of computer code (Ivanova et al., 2020; Liu et al., 2020), providing further evidence that language regions are highly specific to language processing.

Observed dialogue is another interesting boundary case for probing the scope of language regions because listening to dialogue requires the same linguistic processes as listening to monologue, plus additional processes involved in tracking the alternations between speakers. This additional processing may or may not rely on canonical language regions. For instance, compared to processing linguistic input from a single speaker, understanding overheard dialogue requires tracking the difference between at least two speakers' perspectives; thus, understanding dialogue may rely more on Theory of Mind (ToM) - our ability to reason about others' minds - than understanding monologue. ToM selectivity engages a network of regions in right and left temporoparietal junction (RTPJ, LTPJ), middle, ventral, and dorsal parts of medial prefrontal cortex (MMPFC, VMPFC, DMPFC), and precuneus (PC) (Dufour et al., 2013; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). Activity in networks for language and ToM is correlated when processing socially relevant language (Paunov et al., 2019), but these networks still maintain their functional specificity and distribution of roles (Paunov et al., 2022). When comprehending dialogue, tracking shifts in speaker perspective may also specifically be relevant for language processing, as the speaker's perspective can impact the interpretation of the utterance and the predictability of the subsequent response.

Prior research has looked at the neural correlates of comprehension in dialogue when the meaning of an utterance depends on the preceding utterance and contextual information. For example, the utterance "it's hard to give a good presentation" could be a direct response to the question "how difficult is it to prepare a presentation?" (difficult), or an indirect response to the question "what did you think of my presentation?" (not so great; examples adapted from (Bašnáková et al., 2014)). In the brain, ToM network regions including DMPFC and RTPJ, as well as bilateral IFG, and right MTG, responded more to the same utterance when it was an indirect response than when it was a direct responses (Bašnáková et al., 2014; Feng et al., 2017). Another study found that left temporal and frontal regions responded more to indirect than direct replies in question-response pairs (Jang et al., 2013), but the sentences were not controlled for linguistic features between conditions, unlike Bašnáková et al and Feng et al. Individuals with high communicative skills also showed more activation than individuals with low communicative skills for indirect>direct responses in dialogue in regions outside either language or ToM network (Bendtz et al., 2022). These results suggest that the processing of implied meaning in indirect responses mostly occurs outside of the core language network. However, this conclusion remains uncertain. Activation near IFG might imply modulation of the core language network, because part of left IFG is in the canonical language network. This pattern could also reflect activation of nearby 'multiple demand' regions that respond to task difficulty (Blank et al., 2014; Fedorenko et al., 2012; Fedorenko & Blank, 2020), as the indirect replies elicited slower reaction times than the direct replies (Feng et al., 2017) and these studies did not use subject-specific functional regions of interest (ss-fROIs). As experimental stimuli, auditory question-response pairs are well controlled, but afford limited opportunity to recognize, mark and then resolve differences of perspectives between speakers.

Thus, it remains an open question whether the processes that enable an observer to track the alternating perspectives between interlocutors, which are integral to dialogue comprehension, lie within the scope of canonical language regions. To probe the sensitivity of canonical language regions to dialogue, we first directly compared activity in individuals' canonical language regions during naturalistic, audiovisual excerpts of monologue and dialogue. We chose videos in order to maximize meaningful tracking of speakers as individuals with distinct perspectives; consequently, we also needed to create a control condition that accounted for the visual differences between dialogue and monologue. To do this, we used a block-design task ("Sesame Street-Blocked Language" or SS-BlockedLang) with conditions controlling for both comprehensible speech and conversational interaction: videos of two characters (from *Sesame Street*) engaging in either a dialogue or monologues, with the audio for each utterance played normally or temporally reversed (**Experiment 1**). We expected language regions to respond more to comprehensible than incomprehensible speech, but critically, we tested whether language regions responded differently to dialogue versus monologue. A difference in activity could go in either direction: because dialogue requires tracking multiple speakers' speech streams, activity could be higher in dialogue than monologue; however, behaviorally, dialogue is easier to understand than monologue (Fox Tree, 1999; Garrod & Pickering, 2004), and thus may evoke less activity in language regions. To identify language regions in each individual, we used an independent functional localizer task (Scott et al., 2017).

While Experiment 1 was designed to detect differences in experimenter-manipulated aspects of the language clips (dialogue versus monologue, comprehensible versus incomprehensible), one

concern with only using an experimentally-controlled block design is that it might be insensitive to unspecified sources of continuous variation in the neural response (Hamilton & Huth, 2020; Hasson et al., 2004; Hasson & Honey, 2012; Nastase et al., 2020). Complementary efforts within the field of neuroscience have used naturalistic stimuli that are more ecologically valid, such as movies or narratives (Grall & Finn, 2022; Hamilton & Huth, 2020; Nastase et al., 2021; Sonkusare et al., 2019). Therefore, we also designed a complementary second experiment ("Sesame Street-Interleaved Dialogue" or SS-IntDialog) involving longer (1-3 minute) continuous clips of dialogue between two characters (**Experiment 2**). Rather than manipulating forwards vs backwards speech per video clip, as in Experiment 1, in these longer clips, we reversed one of the two character's utterances, such that one character spoke forwards and the other spoke backwards. The visual input was the same for all participants, but the auditory input was not: which character spoke in forward versus backwards speech, in each video, was flipped for half of the participants. First, we tested whether language regions still responded robustly to comprehensible speech, even though the surrounding language (e.g., the speech stream of the other interlocutor) was rendered incomprehensible, meaning that the comprehensible utterances were highly variable and sometimes quite short. Next, we used inter-subject correlation (ISC) analysis to extract stimulus-driven variation in the naturalistic conversation clips, specifically to determine whether language regions only tracked the language-driven variation, or whether they also tracked other aspects of the dialogue that were conveyed independent of which character was comprehensible in the clips (e.g., that two characters alternated speaking, that two characters were present, what objects were in the scene, what the scene was about, etc.). This allowed us to compare the timecourses of response in participants who heard the same version of the stimuli (e.g., the same characters forward vs. backward) to participants who heard the flipped version of

the stimuli (e.g., opposite characters forward vs. backward). If language regions are engaged only in the processing of comprehensible language, and not are not involved in processing any other information that is conveyed in the dialogue clips, then we would expect no correlation in responses across participants who heard flipped versions of the same video.

Regardless of whether canonical language regions differentiate between dialogue and monologue, other regions may show a sensitivity to dialogue. We tested two specific possibilities using individually-defined ss-fROIs: ToM regions and right hemisphere homologues of language regions. Given that speaker alternations in dialogue require integrating across distinct perspectives, we hypothesized that ToM regions might respond differently to dialogue than monologue. Right hemisphere damage can make it more difficult for individuals to make inferences from discourse (Beeman, 1993), and prior work has demonstrated the right hemisphere's preferential involvement in social and contextual aspects of language processing (Friederici, 2011; Frühholz et al., 2012; Ross & Monnot, 2008; Seydell-Greenwald et al., 2020); thus, it was also possible that right hemisphere homologues of language regions might be sensitive to features of dialogue other than just comprehensibility of language. A third possibility is that other regions may be specifically involved in processing comprehensible dialogue, such as regions involved in processing social interactions (Isik et al., 2017); thus, we also performed a whole-brain analysis.

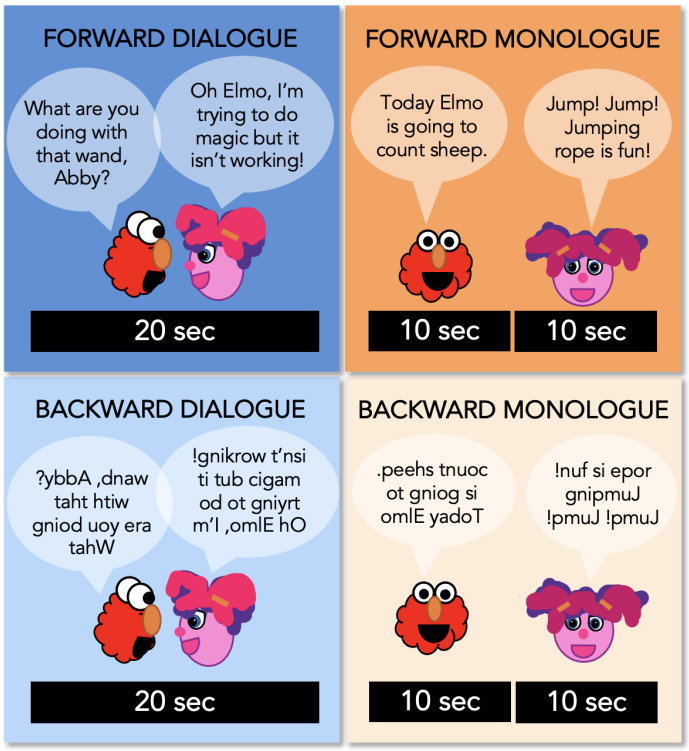
Experiment 1

In Experiment 1, we ran our novel 2x2 block localizer (SS-BlockedLang) along with standard pre-existing localizers for language (Scott et al., 2017) and ToM (Dodell-Feder et al., 2011) to

measure responses to dialogue and monologue speech in subject-specific functional regions of interest for language and ToM.

Methods

Stimuli Design: Our goal was to create a set of stimuli that allowed us to manipulate both comprehensibility and dialogue vs. monologue in a 2x2 block task design (**Figure 1**). To do this, we created a set of 20-second edited audiovisual clips of *Sesame Street* during which either two puppets speak to each other (Dialogue), or a single puppet addresses the viewer (Monologue), while the auditory speech stream is played either normally (Forward) or reversed so as to be incomprehensible (Backward). Dialogue blocks consisted of two characters, both present in the same scene, speaking back-and-forth for a total of 20 seconds, and Monologue blocks consisted of two sequential 10-second clips of a character, present alone. In the Backward conditions, the audio was reversed within each character rather than across the entire clip, ensuring a continuity of voice-character alignment. For instance, in a Backward Dialogue block with Elmo and Abby, Elmo's voice was reversed and played when Elmo was talking, and Abby's voice was reversed and played while Abby was talking.



<u>Figure 1: Experiment 1 Task Design</u>

Participants watched 20-second clips of Dialogue (blue) and Monologue (orange) of Sesame Street, in which the audio is played either Forward or Backward.

We chose to use audiovisual stimuli in order to increase participant engagement with the stimuli, facilitate language comprehension, and emphasize the context of the dialogue by showing multiple characters interacting on the screen. However, using audiovisual stimuli rather than audio-only stimuli introduced a challenge: how do we avoid creating distracting cross-modal mismatches while only varying the auditory, and not the visual, input across conditions? Even infants and young children are sensitive to the congruence between a speaker's mouth movements and the sounds they produce in speech (Gogate & Bahrick, 1998; Lewkowicz & Flom, 2014). To balance these desiderata, we decided to use puppets with rigid mouths (rather

than human actors) so that the congruence between mouth movements and audio would be more similar between the forward and backward speech conditions. Thus, we could counterbalance whether the audio was played forward or backward across participants while showing them the exact same visual clips.

A notable feature of our task is that it uses commercially produced video clips that were not designed for research purposes. Because we intended to eventually use these same stimuli with very young children, video clips were selected from episodes of *Sesame Street* to appeal to a wide age range. The linguistic content is embedded within colorful, dynamic videos with different characters, different voices, and different settings. To retain the natural feel of the clips, the audio was reversed within each utterance of a particular character and carefully overlaid such that the reversed audio still reasonably matched up with the puppets' mouth movements, and each character's "voice" was still unique when the audio was reversed. To create the stimuli, we adhered to the following guidelines: (1) we only selected clips that had an overall neutral or positive valence, (2) we only included clips of puppets, rather than clips with humans and puppets, (3) we excluded clips in which the reversed speech did not align well with mouth movements, and (4) we left non-linguistic sounds in the clips, aiming to retain the integrity of the content. Transcripts and stimuli features can be found here: https://osf.io/whsb7/

Because we selected commercially available clips, we did not control for linguistic properties of the stimuli. Monologue and dialogue blocks did not differ in the number of mental state words per block or the total number of words per block. However, monologue blocks had significantly

longer mean length of utterance, and a lower Flesh-Kincaid reading ease score (see **Supplementary** for details).

Preregistration: Methods and hypotheses were preregistered on OSF: https://osf.io/n4ur5 (validation as language localizer) and https://osf.io/kzdpc (analyses of conversation processing). We did not preregister any hypotheses about right hemisphere homologues of language regions. There were a few deviations from the initial preregistrations for methods: (1) We used a different version of fMRIprep than specified in the preregistration. (2) We decided not to exclude runs of the ToM localizer task based on performance on the true/false questions. (3) We did not preregister testing for effects at the network level, but decided to include these tests along with effects at the ROI level. (4) Not all of the analyses specified in the preregistrations are included in this paper, including: analyses involving laterality index, analyses varying the significance threshold (we used our primary preregistered threshold of $p<.001$), analyses directly comparing the effect of matched vs. mismatched audio in SS-IntDialog in language vs. ToM regions, whole-brain analyses of forward monologue-specific effects, and analyses comparing variance explained in language regions' response using the experimenter-derived regressor of forward/backward speech vs. average timecourse of within-group subjects in the SS-IntDialog task.

Participants: We scanned 20 adults (age: mean(SD) = 23.85(3.70) years, range 18-30 years) who were fluent speakers of English, right-handed, and had no MRI contraindications. Recruitment was restricted to adults with access to the MIT campus according to Covid-19 policies. The protocol was approved by XXX. Informed consent was provided by all participants.

Participants were compensated at a rate of $30/hour for scanning, which is standard for our lab and imaging center.

Experimental Protocol: Data were acquired from a 3-Tesla Siemens Magnetom Prisma scanner located at the XXX using a 32-channel head coil. The scanning session lasted approximately 90 minutes and included a T1 anatomical scan and 10 functional scans: 4 runs of SS-BlockedLang (**Exp. 1**), 2 runs of SS-IntDialog (**Exp. 2**), 2 runs of the auditory language localizer (Scott et al., 2017), and 2 runs of the theory of mind localizer (Dodell-Feder et al., 2011). T1-weighted structural images were acquired in 176 interleaved sagittal slices with 1.0mm isotropic voxels (MPRAGE; TA=5:53; TR=2530.0ms; FOV=256mm; GRAPPA parallel imaging, acceleration factor of 2). Functional data were acquired with a gradient-echo EPI sequence sensitive to Blood Oxygenation Level Dependent (BOLD) contrast in 3mm isotropic voxels in 46 interleaved near-axial slices covering the whole brain (EPI factor=70; TR=2s; TE=30.0ms; flip angle=90 degrees; FOV=210mm). 185 volumes were acquired per run for SS-BlockedLang (TA=6:18), 262 volumes were acquired per run for SS-IntDialog (TA=8:52), 179 volumes were acquired per run for the auditory language localizer (TA=6:06), and 136 volumes were acquired per run for the ToM localizer (TA=4:40). fMRI tasks were run from a MacBook Pro laptop and projected onto a 16”x12” screen. Participants viewed the stimuli through a mirror attached to the head coil. Isocenter to screen + mirror to eye was 42" + 6" for both eyes. SS-BlockedLang and SS-IntDialog tasks were run through PsychoPy3 software version 3.2.4. The auditory language localizer and ToM localizer tasks were run through MATLAB version R2019a and PsychToolbox version 3.0.17.

fMRI Tasks:

SS-BlockedLang Language Task

We used a 2x2 block task design with four conditions: Forward Dialogue, Forward Monologue, Backward Dialogue, and Backward Monologue (**Figure 1**). Participants were asked to watch the 20-second videos and press a button on an in-scanner button box when they saw a still image of Elmo appear on the screen after each 20-second block. Participants completed 4 runs, each 6 min 18 sec long. Each run contained unique clips, and participants never saw a Forward and Backward version of the same clip. Each run contained 3 sets of 4 blocks, one of each condition (total of 12 blocks), with 22-second rest blocks after each set of 4 blocks. Forward and Backward versions of each clip were counterbalanced between participants (randomly assigned Set A or Set B). Run order was randomized for each participant.

Auditory Language Localizer

We used a task previously validated for identifying high-level language processing regions (Scott et al., 2017). Participants listened to Intact and Degraded 18-second blocks of speech. The Intact condition consisted of audio clips of spoken English (e.g., clips from interviews in which one person is speaking), and the Degraded condition consisted of acoustically degraded versions of these clips. Participants viewed a black dot on a white background during the task while passively listening to the auditory stimuli. 14-second fixation blocks (no sound) were present after every 4 speech blocks, as well as at the beginning and end of each run (5 fixation blocks per run). Participants completed two runs, each approximately 6 min 6 sec long. Each run contained 16 blocks of speech (8 intact, 8 degraded).

Theory of Mind Localizer

We used a task previously validated for identifying regions that are involved in ToM and social cognition (Dodell-Feder et al., 2011). Participants read short stories in two conditions: False Beliefs and False Photos. Stories in the False Beliefs condition described scenarios in which a character holds a false belief. Stories in the False Photos condition described outdated photographs and maps. Each story was displayed in white text on a black screen for 10 seconds, followed by a 4-second true/false question based on the story (which participants responded to via the in-scanner button box), followed by 12 seconds of a blank screen (rest). Each run contained 10 blocks. Participants completed two runs, each approximately 4 min 40 sec long.

fMRI Preprocessing and Statistical Modeling: FMRI data were first preprocessed using fMRIPrep 1.2.6 (Esteban et al., 2019), which is based on Nipype 1.1.7 (Gorgolewski et al., 2011). See **Supplementary Materials** for full preprocessing pipeline details. We used a lab-specific script that uses Nipype to combine tools from several different software packages for first-level modeling. Each event regressor was defined as a boxcar convolved with a standard double-gamma HRF, and a high-pass filter (1/128 Hz) was applied to both the data and the model. Artifact detection was performed using Nipype's RapidART toolbox (an implementation of SPM's ART toolbox). Individual TRs were marked as outliers if (1) there was more than .4 units of frame displacement, or (2) the average signal intensity of that volume was more than 3 standard deviations away from the mean average signal intensity. We included one regressor per outlier volume. In addition, we included a summary movement regressor (framewise displacement) and 6 anatomical CompCor regressors to control for the average signal in white matter and CSF. We applied a 6mm smoothing kernel to preprocessed BOLD images. The first-

level model was run using FSL's GLM in MNI space. Subject level modeling was performed with in-lab scripts using Nipype. Specifically, FSL's fixed effects flow was used to combine runs at the level of individual participants. A subject level model was created for each set of usable runs per contrast for each task (up to 4 runs for SS-BlockedLang, and up to 2 runs for SS-IntDialog, LangLoc, and ToMLoc). Runs with more than 20% of timepoints marked as outliers were excluded from analysis (1 run of SS-IntDialog in 1 participant and 1 run of ToMLoc in another participant were excluded for motion). We also excluded 1 run of SS-BlockedLang and 1 run of SS-IntDialog from a participant who reported falling asleep. Output average magnitudes in each voxel in the second level model were then passed to the group level model. Group modeling used in-lab scripts that implemented FSL's RANDOMISE to perform a nonparametric one-sample t-test of the con values against 0 (5000 permutations, MNI space, threshold alpha = .05), accounting for familywise error.

Group Whole Brain Analysis: For group whole brain results, we used a threshold of $p<.001$, corrected via threshold free cluster enhancement (TFCE corrected). For language comprehension, we used the [Forward Dialogue + Forward Monologue] > [Backward Dialogue + Backward Monologue] contrast. To determine whether any other regions in the brain are particularly responsive to comprehensible dialogue, we performed whole-brain analyses using the [Forward Dialogue > Forward Monologue] > [Backward Dialogue > Backward Monologue] contrast. In exploratory analyses, we used an uncorrected threshold of $p<.001$ (two-tailed, 19 degrees of freedom).

Individual Region of Interest Analysis for Language and Theory of Mind: We defined subject-specific functional regions of interest (ss-fROIs) for language as the top 100 voxels activated in an individual, within each of six predefined language search spaces, for the Intact>Degraded contrast using the auditory language localizer task (Fedorenko et al., 2010). The six language search spaces in the left hemisphere included: Left IFGorb, Left IFG, Left MFG, Left AntTemp, Left PostTemp, and Left AngG (similar to Fedorenko et al, 2010; in this case, 6 parcels in left hemisphere were created based on a group-level probabilistic activation overlap map for a sentences>nonwords contrast in 220 adult participants; parcels downloaded from https://evlab.mit.edu/funcloc/). We also looked within the mirror of these search spaces in the right hemisphere (i.e., right hemisphere language homologues). We used the same method as above to define ss-fROIs for theory of mind. In this case, the ToM ss-fROI definition task was the ToM Localizer (Dodell-Feder et al., 2011) using the False Belief > False Photo contrast. The predefined ToM search spaces included 7 regions based on a group random effects analysis with 462 adult participants (Dufour et al., 2013): right and left temporoparietal junction (RTPJ, LTPJ), the precuneus (PC), the dorsal, middle and ventral components of the medial prefrontal cortex (DMPFC, MMPFC and VMPFC), and the right superior temporal sulcus (RSTS). Using the ss-fROIs defined based on the localizer tasks, we then extracted the average magnitude per condition from the SS-BlockedLang task, averaged across all usable runs per participant. Statistical analyses were conducted in R. Conditions were compared using linear mixed effects models; t-tests use Satterthwaite's method. First we tested for network-level fixed effects, with ROI and participants modeled as random effects, using:

lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|ROI)+(1|participantID), REML = FALSE), where b_or_f is backwards or forwards, d_or_m is dialogue or monologue, and ROI is region of

interest within the network. Significance was determined at a level of p<.05 Bonferroni corrected for the 3 networks tested. To test for interactions within individual regions, we used: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE). Significance was determined at a level of p<.05 Bonferroni corrected for the number of ROIs (6 for canonical language regions, 6 for right hemisphere language regions, and 7 for ToM regions).

Individual Regions of Interest Analysis for Comprehensible Conversation: Based on the group whole brain analysis, we performed exploratory analyses in conversation regions of interest, i.e. regions that responded most to comprehensible dialogue in the whole-brain interaction. To do this, we extracted significant clusters using the uncorrected p<.001 thresholded group whole brain contrast for [Forward Dialogue > Backward Dialogue] > [Forward Monologue > Backward Monologue], for clusters with at least 10 voxels. We created 10mm spheres around the center of gravity (COG) for each cluster. To create ss-fROIs, an in-lab script iteratively used the z-stat image of each 3/4 combined runs (i.e., each 'fold') to determine the top 100 voxels for a given subject, ROI, and contrast (in this case, the dialogue interaction contrast). We then used the cope image from the left-out run of a given iteration to extract the betas per condition from these selected top voxels. Statistical analyses were conducted in R. Conditions were compared using linear mixed effects models; t-tests use Satterthwaite's method. To test for interactions within regions, we used: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE).

Overlap Analysis: To determine whether the SS-BlockedLang localizer recruited canonical language regions for the Forward>Backward contrast, we compared responses to the

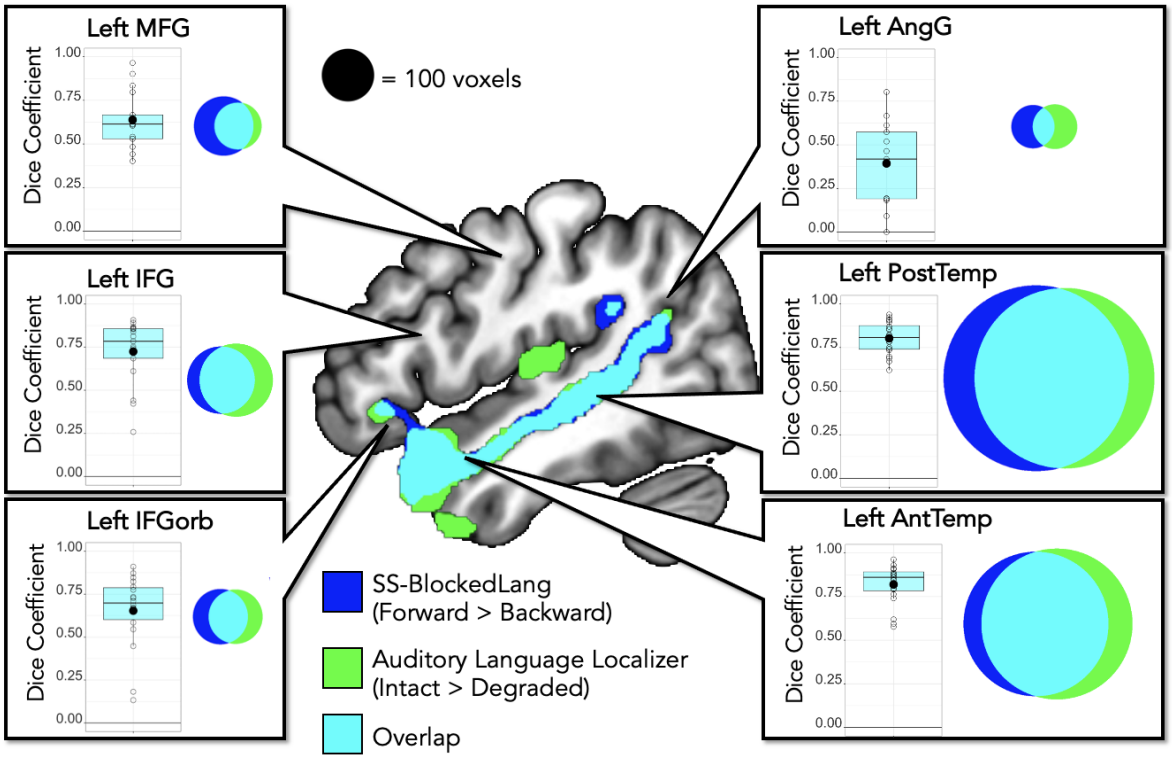
Intact>Degraded contrast from the auditory language localizer. First, across all subjects, we quantified the overlap at the whole-brain level using the group random effects analysis results. Specifically, we calculated Dice coefficients of similarity to capture the extent of overlap in thresholded activation maps, using the formula: Dice coefficient = $2*V_{overlap}/(V_1+V_2)$, where $V_{overlap}$ refers to the number of supra-threshold voxels identified in both tasks, V_1 and V_2 refer to the number of supra-threshold voxels for each of the two tasks, respectively (Rombouts et al, 1997; Wilson et al, 2017). We used a threshold of p<.001 (TFCE corrected) and cluster threshold of k>=10 voxels. Dice coefficients can be described as: low (0-.19), low-moderate (.2-.39), moderate (.4-.59), high-moderate (.6-.79), and high (.8-1) (Wilson et al., 2017). We also calculated the overlap at the whole brain level for each individual subject using a threshold of p<.001. Just as we expected overlap across the whole brain, we also expected a high degree of overlap between the Forward>Backward and Intact>Degraded contrasts within each language region, per subject. We calculated and report Dice coefficients to capture the extent of overlap within each language search space, using a threshold of p<.001 and cluster threshold of k>= 10 voxels to identify suprathreshold voxels. Finally, we calculated the overlap in the top 100 language-selective voxels (e.g., our definition of an ss-fROI) in each subject using the LIT task and using the auditory language localizer task, as a measure of overlap in how these tasks would define language ss-fROIs.

Results

Overlap between tasks

The contrast of Forward versus Backward speech in the *Sesame Street* clips robustly activated the same cortical regions as the contrast of Intact versus Degraded speech using the classic

auditory language localizer, both at the group level (Dice coefficient=.71; high-moderate overlap) and at the level of individual subjects, across the whole brain and within language parcels (**Figure 2**; **Supplementary Tables 1-2, 6**). The mean overlap for all participants was classified as high in left AntTemp and PostTemp regions, high-moderate in left IFGorb, IFG, and MFG, and low-moderate in left AngG.



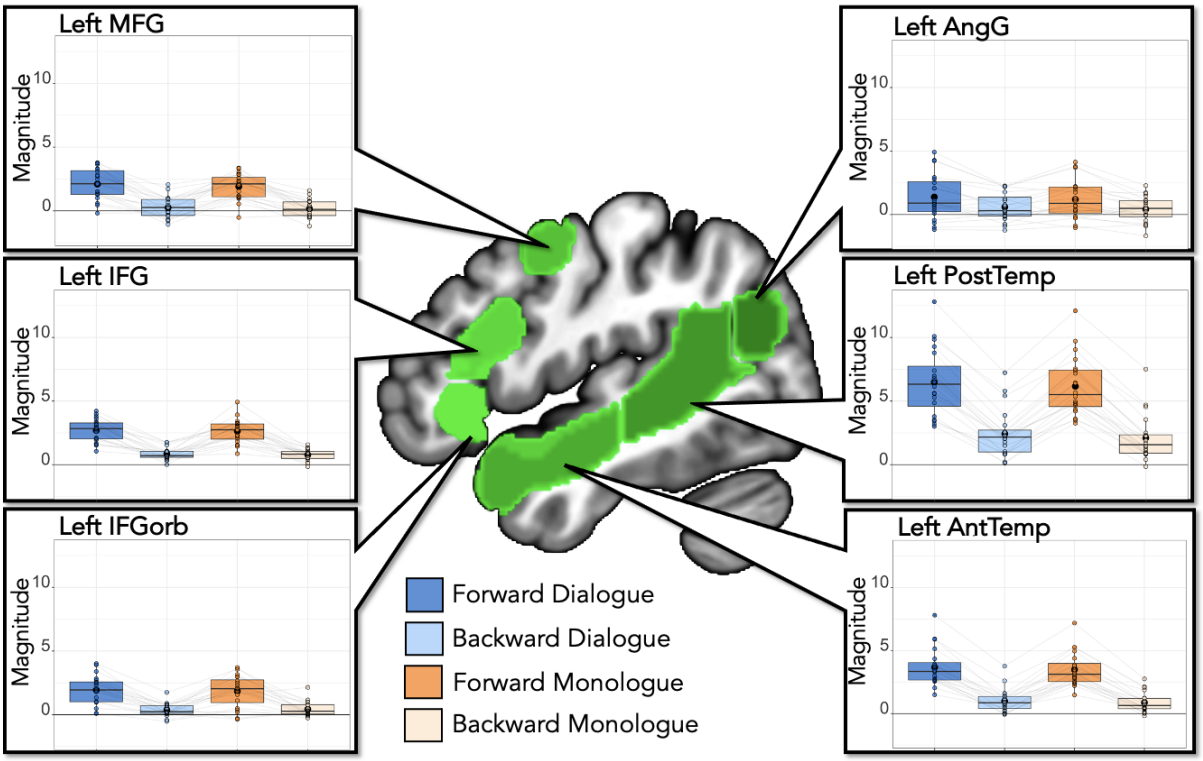
<u>Figure 2: Spatial overlap between SS-BlockedLang and Auditory Language Localizer for language contrast</u>

Center: *Whole brain group random effects analysis for SS-BlockedLang (Forward > Backward; blue) and the Auditory Language Localizer (Intact > Degraded; green) shows high overlap (cyan). Threshold: p<.001, TFCE corrected.* ***Panels:*** *Boxplots show the Dice coefficient (overlap) within each parcel for the language contrast using SS-BlockedLang and the Auditory Language Localizer. Threshold z>3.09; mean shown as filled black circle; open circles*

represent individual participants. Participants without any suprathreshold voxels in at least one contrast for a given parcel do not have a Dice coefficient and are not shown: NA = 2 for IFGorb, NA = 3 for IFG, NA = 3 for MFG, NA = 7 for AngG. Venn diagrams represent the mean number of suprathreshold voxels per contrast, per language parcel (blue: SS-BlockedLang, green: Auditory Language Localizer, cyan: overlap). Mean overlap for participants was in the high-moderate to high range for all regions except left AngG which was low-moderate overlap.

Univariate response to task conditions in language regions

Canonical language network, including all six left-hemisphere language regions defined by the independent auditory language localizer (Scott et al., 2017), showed robust responses to both Forward speech conditions of SS-BlockedLang compared to both Backward speech conditions, as expected (Backward>Forward: Est.=-2.12, S.E.=.15, t-value=-13.69, corrected p-value<.001). This pattern held within each individual ss-fROI (**Figure 3**; **Table 1**; corrected p-values<.001 in every region). There was no main effect of Dialogue compared to Monologue in canonical language network (Dialogue>Monologue: Est.=.17, S.E.=.15, t-value=1.13, corrected p-value=.78), nor an interaction between comprehensibility and dialogue (Backward>Forward*Dialogue>Monologue: Est.=-.05, S.E.=.22, t-value=-.22, corrected p-value=1; individual language ss-fROI results in **Figure 3; Table 1**; corrected p-values>.1 in every region for dialogue and interaction).



<u>Figure 3: SS-BlockedLang average magnitude by condition within language regions</u>

Center: *Left hemisphere language parcels overlaid on template brain (green; parcels include left IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/).*

Panels: *Average response magnitude (betas) per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for language (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. There was a main effect of Forward speech compared to Backward speech, but no effect of Dialogue speech compared to Monologue speech within language regions.*

ROI	Backward v.	Dialogue v. Monologue	Interaction

	Forward		
Left IFGorb	Est.= -1.57 S.E.= 0.20 t-value= -7.69 **p< .001 ***	Est.= 0.10 S.E.= 0.20 t-value=0.49 p= 0.63	Est.= -0.15 S.E.= 0.29 t-value= -0.51 p= 0.61
Left IFG	Est.= -1.82 S.E.= 0.16 t-value= -11.23 **p< .001 ***	Est.= 0.06 S.E.= 0.16 t-value= 0.35 p= 0.73	Est.= 0.04 S.E.= 0.23 t-value= 0.19 p=0.85
Left MFG	Est.= -1.84 S.E.= 0.19 t-value= -9.77 **p< .001 ***	Est.=0.19 S.E.= 0.19 t-value= 1.00 p= 0.32	Est.= -0.10 S.E.= 0.27 t-value= -0.39 p= 0.70
Left AntTemp	Est.= -2.68 S.E.= 0.16 t-value= -16.26 **p< .001 ***	Est.= 0.19 S.E.= 0.16 t-value= 1.13 p= 0.26	Est.= -0.07 S.E.= 0.23 t-value= -0.29 p= 0.78
Left PostTemp	Est.= -4.06 S.E.= 0.21 t-value= -19.12 **p< .001 ***	Est.= 0.33 S.E.= 0.21 t-value= 1.56 p= 0.12	Est.= -0.01 S.E.= 0.30 t-value= -0.03 p= 0.98
Left AngG	Est.= -0.79 S.E.= 0.19 t-value= -4.25 **p< .001 ***	Est.= 0.19 S.E.= 0.19 t-value= 1.01 p= 0.32	Est.= -0.01 S.E.= 0.26 t-value= -0.05 p= 0.96

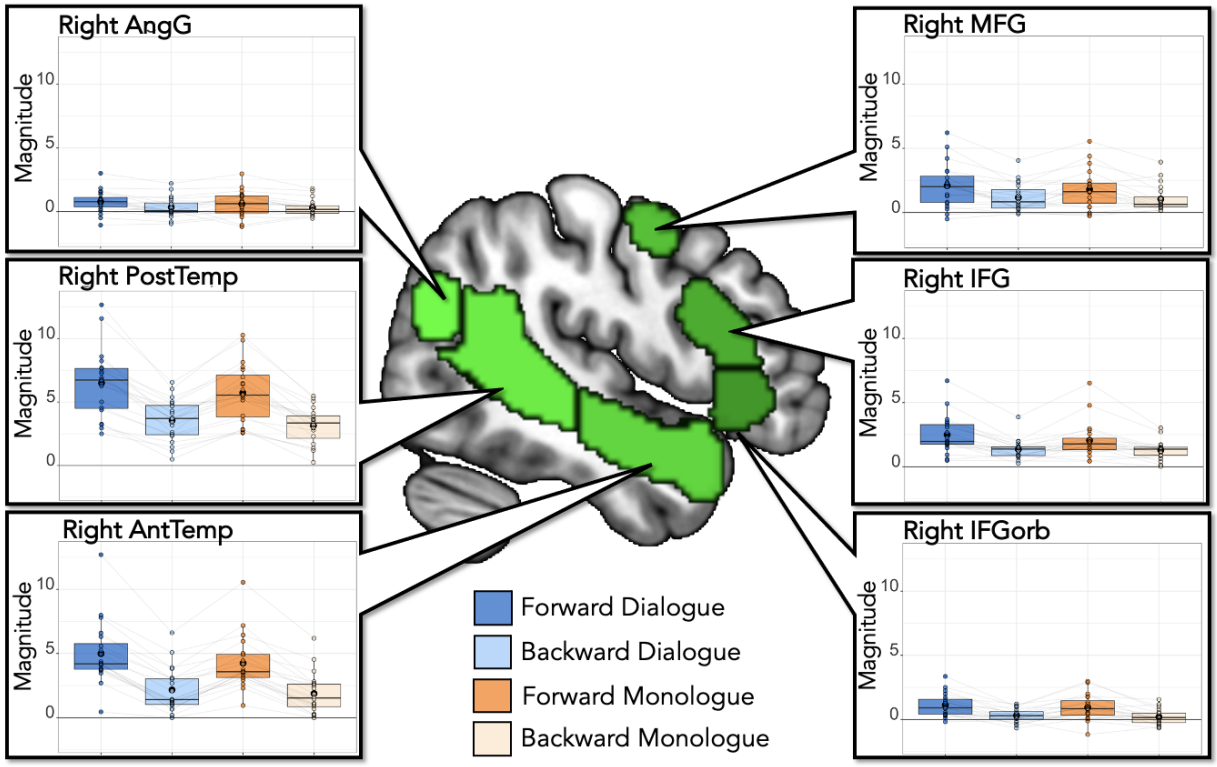
Table 1: SS-BlockedLang statistics in language regions

*Within each language ss-fROI, there was a significant difference between Forward and Backward speech, but no difference between Monologue and Dialogue, and no interaction. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*

** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

Univariate response to task conditions outside canonical language regions

There were effects of dialogue in regions of cortex outside the canonical left-hemisphere language network. First, we examined right hemisphere homologues of language regions, which responded more to forward than backward speech (Backward>Forward: Est.=-1.51, S.E.=.17, t-value=-9.05, corrected p-value<.001), and more to dialogue than monologue speech (Dialogue>Monologue: Est.=.44, S.E.=.17, t-value=2.61, corrected p-value=.028), though showed no interaction between comprehensibility and dialogue (Backward>Forward*Dialogue>Monologue: Est.=-.26, S.E.=.24, t-value=-1.11, corrected p-value=.80). Individually, all of these regions responded more to forward than backward speech, and AntTemp and PostTemp responded more to dialogue than monologue (**Figure 4; Table 2**); there were no significant interactions between comprehensibility (forward/backwards) and dialogue (dialogue/monologue) in any individual regions.



<u>Figure 4: SS-BlockedLang average magnitude by condition within right homologue language regions</u>

Center: *Right hemisphere language parcels (mirror of left hemisphere parcels) overlaid on template brain (green; parcels include right IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/).* ***Panels***: *Average response magnitude (betas) per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for right language homologues (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. There was a main effect of Forward speech compared to Backward speech in all regions, and a main effect of Dialogue speech compared to Monologue speech in right AntTemp and PostTemp.*

ROI	Backward v. Forward	Dialogue v. Monologue	Interaction
Right IFGorb	Est.= -0.81 S.E.= 0.15 t-value= -5.44 **p< .001 ***	Est.= 0.19 S.E.= 0.15 t-value= 1.26 p= 0.21	Est.= -0.07 S.E.= 0.21 t-value= -0.32 p= 0.75
Right IFG	Est.= -1.14 S.E.= 0.18 t-value= -6.46 **p< .001 ***	Est.= 0.44 S.E.= 0.18 t-value= 2.49 p= 0.02	Est.= -0.38 S.E.= 0.25 t-value= -1.51 p= 0.14
Right MFG	Est.= -0.92 S.E.= 0.19 t-value= -4.81 **p< .001 ***	Est.= 0.29 S.E.= 0.19 t-value= 1.53 p= 0.13	Est.= -0.17 S.E.= 0.27 t-value= -0.62 p= 0.54
Right AntTemp	Est.= -2.83 S.E.= 0.20 t-value= -14.31 **p< .001 ***	Est.= 0.74 S.E.= 0.20 t-value= 3.76 **p< .001 ***	Est.= -0.46 S.E.= 0.28 t-value= -1.63 p= 0.11
Right PostTemp	Est.= -2.95 S.E.= 0.26 t-value= -11.55 **p< .001 ***	Est.= 0.77 S.E.= 0.26 t-value= 3.03 **p= 0.004 ***	Est.= -0.36 S.E.= 0.36 t-value= -1.00 p= 0.32
Right AngG	Est.= -0.43 S.E.= 0.12 t-value= -3.62 **p< .001 ***	Est.= 0.18 S.E.= 0.12 t-value= 1.49 p= 0.14	Est.= -0.14 S.E.= 0.17 t-value= -0.84 p= 0.40

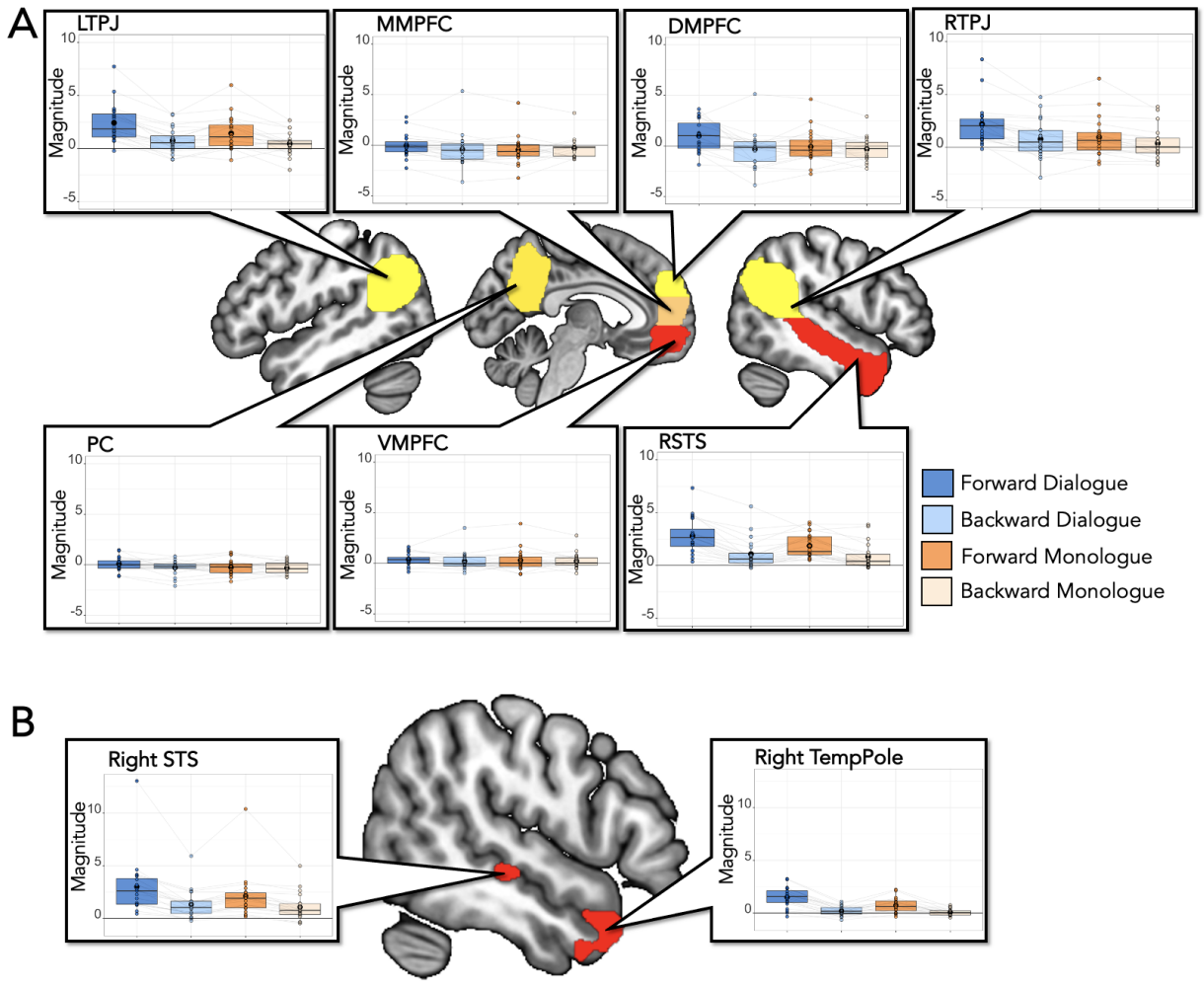
Table 2: SS-BlockedLang statistics in right hemisphere language region homologues

There was a significant difference between Forward and Backward speech within each right language homologue ss-fROI, and a main effect of Dialogue speech compared to Monologue speech in AntTemp and PostTemp, but no interaction. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model:

*lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*

** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

Next, we examined responses to each task condition in ToM regions. ToM network responded more to forward than backward speech (Backward>Forward: Est.=-1.00, S.E.=.15, t-value=-6.80, corrected p-value<.001), more to dialogue than monologue (Dialogue>Monologue: Est.=.73, S.E.=.15, t-value=4.93, corrected p-value<.001), and showed an interaction between comprehensibility and dialogue (Backward>Forward*Dialogue>Monologue: Est.=-.61, S.E.=.21, t-value=-2.90, corrected p-value=.012). Individually, four out of 7 regions responded more to forward than backward speech, and more to dialogue than monologue (DMPFC, LTPJ, RTPJ, and RSTS; **Figure 5a; Table 3**). DMPFC and RTPJ had a significant interaction between comprehensibility and dialogue, responding most for Forward Dialogue.



<u>Figure 5: SS-BlockedLang whole brain interaction for comprehensible dialogue</u>

(A) Center*: Theory of mind parcels overlaid on template brain (parcels include LTPJ, MMPFC, DMPFC, RTPJ, PC, VMPFC, and RSTS from* (Dufour et al., 2013)*). **Panels**: Average response magnitude per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for theory of mind (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. There was a main effect of Forward compared to Backward*

speech and a main effect of Dialogue compared to Monologue in DMPFC, LTPJ, RTPJ, and RSTS, and an interaction in DMPFC and RTPJ.

***(B) Center**: Right superior temporal sulcus and right temporal pole were activated for [Forward Dialogue > Forward Monologue] > [Backward Dialogue > Backward Monologue]. Threshold p<.001, uncorrected (df=19, two-tailed). Nothing survives at TFCE corrected threshold. Shown here are clusters for Right STS and Right Temporal Pole; not shown: clusters in Left STS and Left Cerebellum. **Panels**: Average response magnitude per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for conversation, based on spheres around center of gravity voxels from the group whole-brain interaction contrast (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. Results shown for Right STS and Right Temporal Pole. There was a main effect of Forward>Backward and a main effect of Dialogue>Monologue in these regions, as well as an interaction in Right Temporal Pole.*

ROI	Backward v. Forward	Dialogue v. Monologue	Interaction
DMPFC	Est.= -1.32 S.E.= 0.27 t-value= -4.86 **p< .001 ***	Est.= 1.07 S.E.= 0.27 t-value= 3.92 **p< .001 ***	Est.= -1.08 S.E.= 0.38 t-value= -2.81 **p= 0.007 ***
MMPFC	Est.= -0.36 S.E.= 0.20 t-value= -1.84 p= 0.07	Est.= 0.42 S.E.= 0.20 t-value= 2.11 p= 0.04	Est.= -0.52 S.E.= 0.28 t-value= -1.87 p= 0.07
VMPFC	Est.= -0.17	Est.= 0.08	Est.= -0.10

	S.E.= 0.14 t-value= -1.20 p= 0.24	S.E.= 0.14 t-value= 0.58 p= 0.56	S.E.= 0.20 t-value= -0.49 p= 0.62
LTPJ	Est.= -1.66 S.E.= 0.22 t-value= -7.73 **p< .001 ***	Est.= 0.99 S.E.= 0.22 t-value= 4.61 **p< .001 ***	Est.= -0.69 S.E.= 0.30 t-value= -2.27 p= 0.03
PC	Est.= -0.37 S.E.= 0.14 t-value= -2.74 p= 0.008	Est.= 0.35 S.E.= 0.14 t-value= 2.56 p= 0.01	Est.= -0.30 S.E.= 0.19 t-value= -1.59 p= 0.12
RTPJ	Est.= -1.42 S.E.= 0.19 t-value= -7.37 **p< .001 ***	Est.= 1.25 S.E.= 0.19 t-value= 6.47 **p< .001 ***	Est.= -0.85 S.E.= 0.27 t-value= -3.12 **p= 0.003 ***
RSTS	Est.= -1.72 S.E.= 0.19 t-value= -9.21 **p< .001 ***	Est.= 0.94 S.E.= 0.19 t-value= 5.06 **p< .001 ***	Est.= -0.69 S.E.= 0.26 t-value= -2.62 p= 0.01

Table 3: SS-BlockedLang statistics in theory of mind regions

Within ToM ss-fROIs, there was a main effect of Forward compared to Backward speech and a main effect of Dialogue compared to Monologue in DMPFC, LTPJ, RTPJ, and RSTS, and an interaction in DMPFC and RTPJ. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model:

*lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*

** indicates significance level p<.05, Bonferroni corrected for 7 ROIs (p<.0071)*

Finally, to empirically test for regions that specifically respond to comprehensible dialogue, we performed a whole brain analysis for the following interaction: [Forward Dialogue>Forward Monologue]>[Backward Dialogue>Backward Monologue] (**Figure 5b**). Four clusters were identified using an uncorrected threshold of p<.001 in the right temporal pole, right STS, left

STS, and left cerebellum (none survived TFCE correction for multiple comparisons). In exploratory analyses, we extracted activity in individual participants in individually defined ss-fROIs (within the 10mm sphere search spaces around center of gravity coordinates from the group results), using a leave-one-run-out approach (**Figure 5b; Table 4**). All four regions responded more to Dialogue than Monologue, and all except left Cerebellum responded more to Forward than Backward speech. There was an interaction between comprehensibility and dialogue in the right Temporal Pole and left Cerebellum.

ROI	Voxels	MAX T-value	Peak X,Y,Z (mm)	COG X,Y,Z (mm)	Backward v. Forward	Dialogue v. Monologue	Interaction
Right Temporal Pole	299	6.64	54, 18, -26	52, 12.6, -30.9	Est.= -0.65 S.E.= 0.13 t-value= -4.95	Est.= 0.77 S.E.= 0.13 t-value= 5.82	Est.= -0.64 S.E.= 0.19 t-value= -3.41
Right STS	98	6.33	50, -26, -6	52, -23.9, -6.18	Est.= -1.06 S.E.= 0.25 t-value= -4.27	Est.= 0.87 S.E.= 0.25 t-value= 3.53	Est.= -0.62 S.E.= 0.35 t-value= -1.79
Left Crus 2 (Cerebellum)	29	4.61	-26, -78, -34	-28.1, -80, -34.6	Est.= -0.18 S.E.= 0.13 t-value= -1.43	Est.= 0.46 S.E.= 0.13 t-value= 3.59	Est.= -0.48 S.E.= 0.18 t-value= -2.65
Left STS	14	4.74	-50, -30, -4	-50.3, -30.6, -3.42	Est.= -1.20 S.E.= 0.17 t-value= -7.08	Est.= 0.54 S.E.= 0.17 t-value= 3.18	Est.= -0.46 S.E.= 0.24 t-value= -1.92

Table 4: SS-BlockedLang comprehensible dialogue regions

Significant clusters at p<.001 (uncorrected, df=19, two-tailed). Peak coordinates and center of gravity (COG) for the cluster (weighted average of the coordinates by the intensities within the cluster). No significant voxels at p<.001 TFCE corrected. Within ss-fROIs defined based on 10mm spheres around the group cluster COG coordinates, there was a higher response to Forward than Backward speech in all regions except Left Cerebellum, and a main effect of Dialogue speech compared to Monologue speech in all regions, as well as an interaction in

*Right Temporal Pole and Left Cerebellum. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE). P-values not reported since analyses were exploratory.*

Summary

These results suggest that dialogue does not modulate language processing in canonical left hemisphere cortical language regions. The magnitude of response in classic language regions appears to be determined by the presence of local structure in linguistic stimuli (common to both Forward conditions), while distinct cortical regions are sensitive to the differences between dialogue and monologue speech, including some theory of mind regions (DMPFC, LTPJ, RTPJ, and RSTS) and right language homologues (right AntTemp and PostTemp), as well as other regions identified by exploratory whole-brain analyses (in right Temporal Pole, right STS, left Crus II in cerebellum, and left STS).

Next, we decided to further probe the sensitivity of language regions to features of dialogue by using longer clips of dialogue with interleaved forward and backward speech in **Experiment 2**. Rather than blocks of all-forward and all-backward speech, one character's audio stream was played forward, while the other character's audio stream was played backward (which character was forward versus backward was counterbalanced between participants). This approach complements **Experiment 1** in a few ways. First, we measured canonical language regions' responses to forward speech within the temporal structure of natural dialogue, i.e. frequent short utterances, instead of long blocks. Second, the longer dialogue clips in **Experiment 2** allowed us

to use ISC analyses to directly measure the influence of linguistic structure, compared to all other visual and abstract semantic structure of the dialogue, on the timecourse of activity in canonical language regions.

Experiment 2: SS-IntDialog

In Experiment 2, we ran our novel localizer with 1-3 minute segments of dialogue, with forward and backward speech alternating by character (SS-IntDialog), to measure responses to dialogue over time in subject-specific functional regions of interest for language and ToM.

Methods

Participants, experimental protocol, and fMRI tasks, and fMRI processing are identical to Experiment 1. In the same session, we also had participants undergo a second novel fMRI task (**SS-IntDialog**) that used clips of longer conversational interactions.

Stimuli Design: General methods for stimuli design were similar to **Experiment 1**. Our goal was to create a set of stimuli that were more naturalistic than the first task, but still allowed us to introduce an experimental manipulation of forward compared to backward speech. To do this, we selected full scenes of dialogue from *Sesame Street* during which two puppets speak to each other (the selected scenes ranged from 1-3 minutes, and we played the entire scene). Like the clips used in the SS-BlockedLang task, these scenes varied in terms of their visual properties (such as objects and setting), topic, and characters. For each clip, we reversed the audio for one character's utterances, but left the other character's audio forward (**Figure 6A**). We had two versions of each clip, such that one group of participants heard one character forward (e.g., Elmo

forward and Abby backward) and the other group of participants heard the other character forward (e.g., Abby forward and Elmo backward). This allowed us to calculate ISCs between a held-out subject's timecourse and (1) the average timecourse for other participants who heard the same version of the videos, and (2) the average timecourse for the participants who heard the opposite version of the videos, within ss-fROIs (**Figure 6B**). Comprehensible utterances varied in length from .46 to 34.68 seconds, with a mean(SD) of 3.74(3.84) seconds (**Figure 6B**).

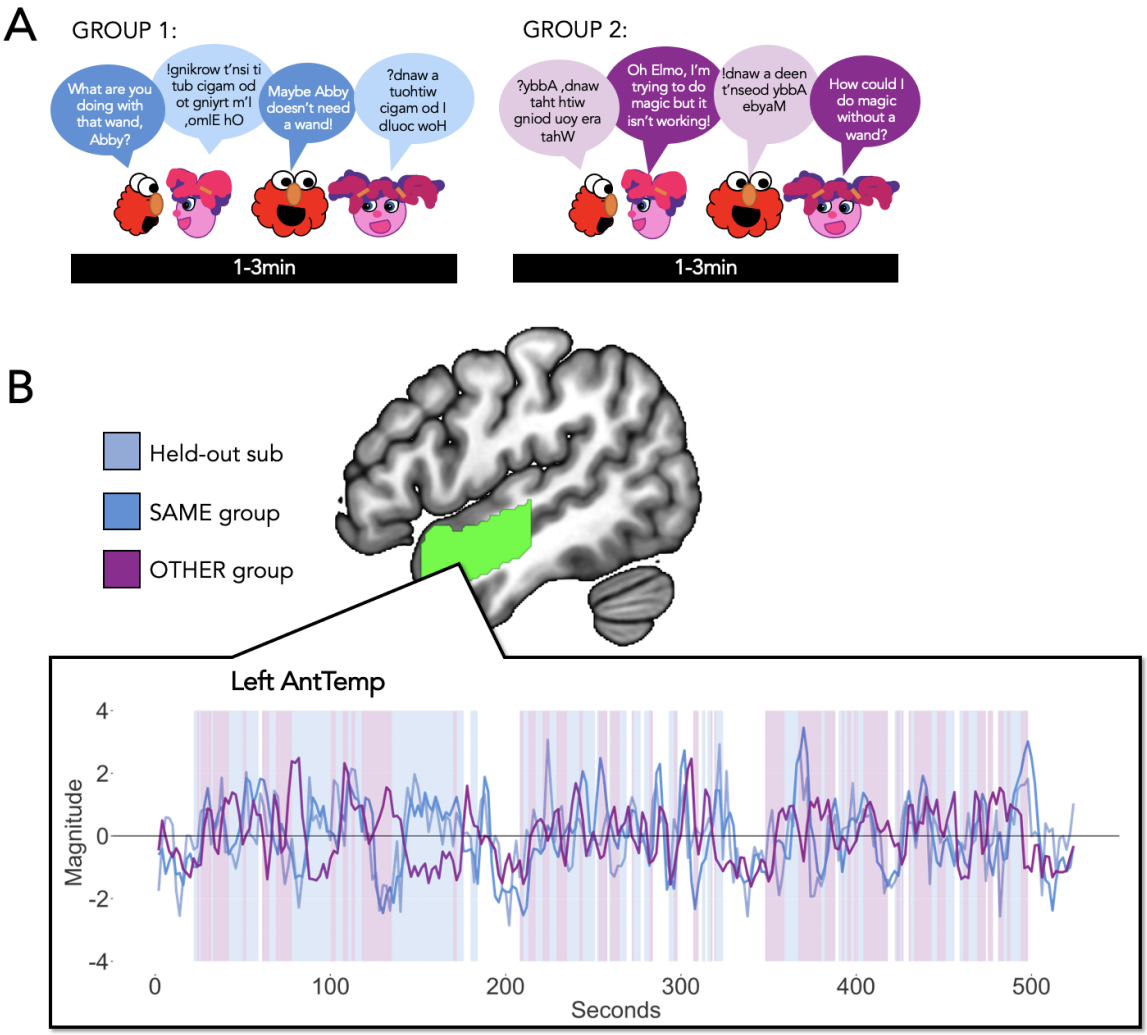


Figure 6: SS-IntDialog Task Design

(A) *Participants watched 1-3 minute clips of Sesame Street in which two characters have a conversation. The audio from one character is played Forward while the second is played Backward. Participants were randomly assigned to hear one of the two versions (with opposite characters played Forward/Backward). Participants watched two runs, each containing 3 clips with 20 seconds of fixation before and after each clip.*

(B) Center*: One language ROI (Left AntTemp, green). ss-fROIs were created per subject within language parcels, theory of mind parcels, and conversation spherical parcels.* ***Within box, left****: Example timecourse for one run of SS-IntDialog, for one participant (light blue), the average of the other participants who heard the exact same version of the run (darker blue), and the average of the participants who heard the opposite version of the run (purple). Background shading indicates when speech is forward (blue) or backward (purple) from the perspective of the held-out participant (opposite for the "other" group: purple is forward and blue is backward).*

SS-IntDialog Language Task

Participants watched 1–3-minute dialogue clips of *Sesame Street* in which one character's audio stream was played Forward and the other was played Backward. Additional sounds in the video (e.g., blowing bubbles, a crash from something falling) were played forwards. Participants watched the videos and pressed a button on an in-scanner button box when they saw a still image of Elmo appear on the screen immediately after each block. Participants completed 2 runs, each approximately 8 min 52 sec long. Each run contained unique clips, and participants never saw a version of the same clip with the Forward/Backward streams reversed. Each run contained 3

clips, 1-3 minutes each, presented in the same order. Between each video, as well as at the beginning and end of the run, there was a 22-second fixation block. Versions of each clip with the opposite character Forward and Backward were counterbalanced between participants (randomly assigned Set A or Set B). 11 participants saw version A, and 9 participants saw version B (1 run from group A was excluded due to participant falling asleep, and one run from group B was excluded due to motion). Run order was randomized for each participant (random sequence 1-2). Transcripts and stimuli features can be found here: https://osf.io/whsb7/

Intersubject Correlation Analysis: For the SS-IntDialog task, each participant saw two runs, each of which contained 3 different video clips (in the same order within a run). Half the participants saw version A, and half of the participants saw version B of these runs (same videos, different audio streams). That is, if Elmo is speaking forward in the first clip in Run 1 version A, Elmo would be speaking backward in the first clip in Run 1 version B. We performed ISC analyses across the entire run, including the rest blocks between clips. ISC analyses were performed using in-lab scripts modeled after the tutorials in https://naturalistic-data.org/ (Chang et al., 2020). The preprocessed data was smoothed with a 6mm kernel, and then denoised using a GLM (6 realignment parameters, their squares, their derivatives, and squared derivatives), with outliers excluded using a dummy code, and average CSF activity and linear and quadratic trends regressed out. The timecourse was z-transformed to be centered at 0.

First, we extracted the timecourse per participant, per run for each language ss-fROI (defined as specified in **Exp. 1**, using the auditory language localizer). Using a leave-one-subject out approach, we calculated the correlation between the held-out subject's timecourse (i.e. the

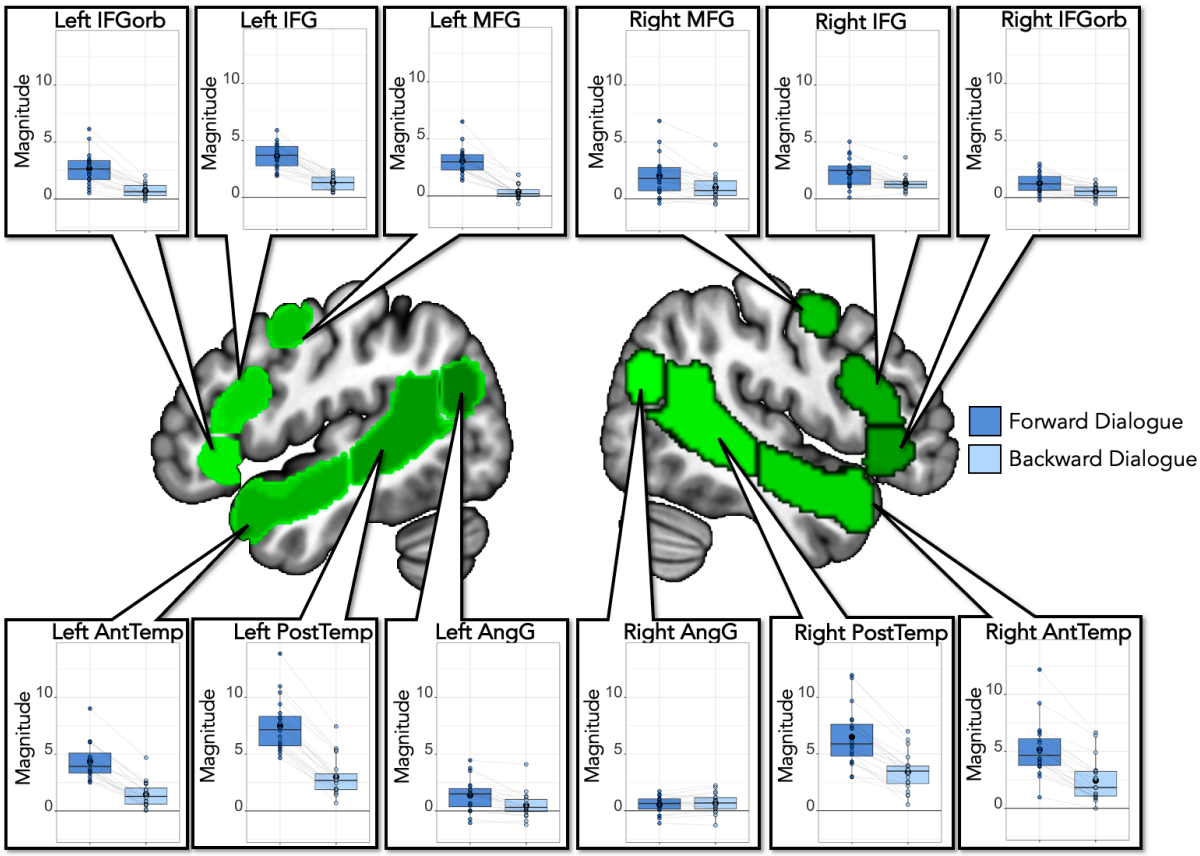
average response of that subject across all 100 voxels in that ROI) and (1) the average timecourse of the remaining participants who watched the same version of the stimuli, and (2) the average timecourse of the participants who watched the opposite version of the stimuli, for each language region. Next, we did the same analyses using the extracted timecourses per participant, per run for each of the ToM ss-fROIs. Finally, we repeated the same analysis with the extracted timecourses per participant, per run for each conversation ss-fROI, defined as the top 100 voxels for the [Forward Dialogue>Forward Monologue]>[Backward Dialogue>Backward Monologue] interaction contrast within 10mm spheres centered at the center of gravity point for each significant cluster in the group map (**Table 4**).

Results

Univariate response to forward and backward speech

By modeling the onset and offset of each utterance within the extended SS-IntDialog dialogues, we replicated the robust response to Forward utterances, and the very low response to Backward utterances, in canonical left-hemisphere language network (Backward>Forward: Est.= -2.54, S.E.= 0.16, t-value=-15.53, corrected p-value<.001), and in individual language regions (**Figure 7**, **Supplementary Table 7**). Right hemisphere homologues of language regions likewise responded more to Forward than Backward speech at a network level (Backward>Forward: Est.= -1.40, S.E.=.19, t-value= -7.46, corrected p-value<.001), and at the level of individual regions with the exception of right AngG (**Figure 7**, **Supplementary Table 7**). Both when looking across the whole brain and looking within language parcels, there was moderate to high-moderate overlap between the Forward>Backward contrast from SS-IntDialog and the Intact>Degraded contrast from the auditory language localizer (**Supplementary Tables 1, 4**).

Thus, even with the linguistic content of the other speaker's utterance removed, canonical language regions still responded more to comprehensible than incomprehensible speech.

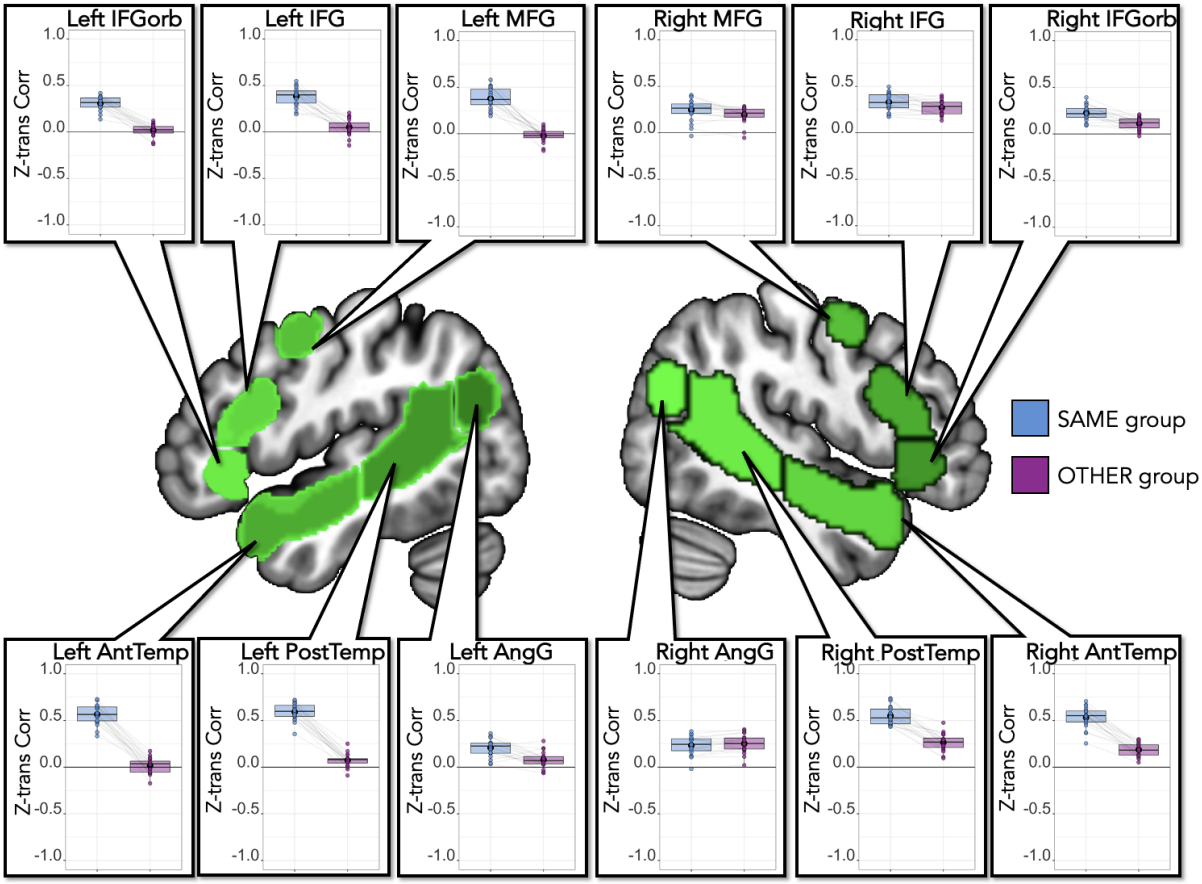


<u>Figure 7: SS-IntDialog average magnitude by condition within language regions and right language regions homologues</u>

Center: *Left hemisphere language parcels and right-hemisphere homologues overlaid on template brain (green; parcels include left and right IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/).* ***Panels***: *Average response magnitude per individual for each condition in the SS-IntDialog task was extracted from subject-specific functional regions of interest for language (blue: Forward Dialogue; light blue: Backward Dialogue). All regions except right AngG responded more to Forward than Backward speech.*

Timecourse of response to dialogue videos

In addition to looking at the magnitude of the response to Forward and Backward speech during dialogue, we also compared the timecourse of the response across participants. The timecourse of response in canonical left hemisphere language regions was correlated across participants who saw the same version of the extended dialogue, with the same character's speech played forward. Thus, even the short and variable utterances within these dialogues could evoke reliable responses, consistently across participants. When comparing the timecourse to participants hearing the opposite character's speech played forward, there was little to no correlation in language regions (**Figure 8**; **Table 5**). This suggests that responses were driven by language comprehensibility, rather than visual and abstract semantic structure of the dialogues preserved between the groups (e.g., the sequence of visual images, the topic of the conversation, etc.).



<u>Figure 8: SS-IntDialog correlations within language regions and right homologues</u>

Center*: Left hemisphere language parcels and right-hemisphere homologues overlaid on template brain (green; parcels include left and right IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/).* ***Panels****: Average z-transformed Pearson's correlation between each held-out subject's timecourse within each ss-fROI and the average timecourse of the remaining participants who viewed and listened to the same version of the stimuli (blue) and the average of the participants who heard the opposite audio stream (purple). Each individual's datapoints are connected by light gray lines. Within-group correlations were higher than between-group correlations in all regions except right IFG and AngG.*

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
Left IFGorb	M(SD) = 0.31(0.07); range = 0.14-0.42 One-sample t-test: t-value = 19.02, **p< .001 ***	M(SD) = 0.02(0.06); range = -0.13-0.12 One-sample t-test: t-value = 1.19, p-value = 0.25	t = 14.32, **p< .001 ***
Left IFG	M(SD) = 0.38(0.10); range = 0.19-0.55 One-sample t-test: t-value = 17.30, **p< .001 ***	M(SD) = .05(0.09); range = -.14-.21 One-sample t-test: t-value = 2.59, p-value = 0.02	t = 10.02, **p< .001 ***
Left MFG	M(SD) = 0.38(0.11); range = 0.19-0.58 One-sample t-test: t-value = 15.07, **p< .001 ***	M(SD) = -0.02(0.07); range = -0.18-0.10 One-sample t-test: t-value = -1.27, p-value = 0.22	t = 11.95, **p< .001 ***
Left AntTemp	M(SD) = 0.57(0.11); range = 0.33-0.73 One-sample t-test: t-value = 24.01, **p< .001 ***	M(SD) = 0.02(0.08); range = -0.17-0.17 One-sample t-test: t-value = 0.81, p-value = 0.43	t = 18.32, **p< .001 ***
Left PostTemp	M(SD) = 0.59(0.09); range = 0.36-0.72 One-sample t-test: t-value = 28.96, **p< .001 ***	M(SD) = 0.07(0.07); range = -0.09-0.25 One-sample t-test: t-value = 4.53, **p< .001 ***	t = 22.36, **p< .001 ***
Left AngG	M(SD) = 0.21(0.09); range = 0.03-0.36 One-sample t-test: t-value = 9.77, **p< .001 ***	M(SD) = 0.08(0.08); range = -0.06-0.28 One-sample t-test: t-value = 4.20, **p< .001 ***	t = 4.73, **p< .001 ***

Table 5: SS-IntDialog timecourse correlations within language regions

Average z-transformed Pearson's correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected p-values reported). Paired t-test shows that there

were higher within-group than between-group correlations for each canonical language region (uncorrected p-values reported).

** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
Right IFGorb	M(SD) = 0.23(0.08); range = 0.09-0.39 One-sample t-test: t-value = 12.44, **p< .001 ***	M(SD) = 0.11(0.07); range = -0.02-0.21 One-sample t-test: t-value = 7.46, **p< .001 ***	t = 5.36, **p< .001 ***
Right IFG	M(SD) = 0.33(0.10); range = 0.17-0.50 One-sample t-test: t-value = 15.39, **p< .001 ***	M(SD) = 0.27(0.08); range = 0.13-0.40 One-sample t-test: t-value = 15.64, **p< .001 ***	t = 2.89, p-value = 0.009
Right MFG	M(SD) = 0.24(0.12); range = -0.03-0.41 One-sample t-test: t-value = 9.41, **p< .001 ***	M(SD) = 0.19(0.09); range = -0.05-0.29 One-sample t-test: t-value = 9.92, **p< .001 ***	t = 3.00, **p-value = 0.007 ***
Right AntTemp	M(SD) = 0.530.11); range = 0.25-0.71 One-sample t-test: t-value = 21.39, **p< .001 ***	M(SD) = 0.19(0.07); range = 0.05-0.30 One-sample t-test: t-value = 11.71, **p< .001 ***	t = 14.01, **p< .001 ***
Right PostTemp	M(SD) = 0.55(0.10); range = 0.43-0.74 One-sample t-test: t-value = 24.37, **p< .001 ***	M(SD) = 0.26(0.09); range = 0.10-0.48 One-sample t-test: t-value = 12.67, **p< .001 ***	t = 10.62, **p< .001 ***
Right AngG	M(SD) = 0.23(0.10); range = -0.02-0.38 One-sample t-test: t-value = 10.25, **p< .001 ***	M(SD) = 0.25(0.10); range = 0.02-0.40 One-sample t-test: t-value = 11.46, **p< .001 ***	t = -1.39, p-value = 0.18

Table 6: SS-IntDialog timecourse correlations within right language region homologues

Average z-transformed Pearson's correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected p-values reported). Paired t-test shows that there were higher within-group than between-group correlations for each right hemisphere language region except right IFG and AngG (uncorrected p-values reported).

** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

The dialogue videos also evoked shared temporal structure across and within participant groups in other cortical regions. In right hemisphere language regions, both the within-group and between-group Pearson's correlations were greater than 0 (**Figure 8; Table 6**), though the within-group correlations were higher than the between-group correlations for all of the regions except right IFG and AngG. Brain regions defined by the independent ToM localizer also showed significant correlations in the timecourse of response with the same and also with opposite videos (with the exception of VMPFC and MMPFC), again with higher within-group correlations than between-group correlations for all regions except MMPFC and PC (**Figure 9A**; **Table 7**). The brain regions in right STS, right temporal pole, left STS, and left cerebellum identified as specifically responsive to comprehensible dialogue in SS-BlockedLang similarly showed correlated timecourses when participants watched matched or opposite videos (with the exception of left STS for the between-groups correlation; **Figure 9B**; **Table 8**), but had higher correlations for the matched videos. Thus, the preserved visual and abstract semantic structure of

the dialogue contributed to reliable cortical responses outside of the canonical left hemisphere language regions.

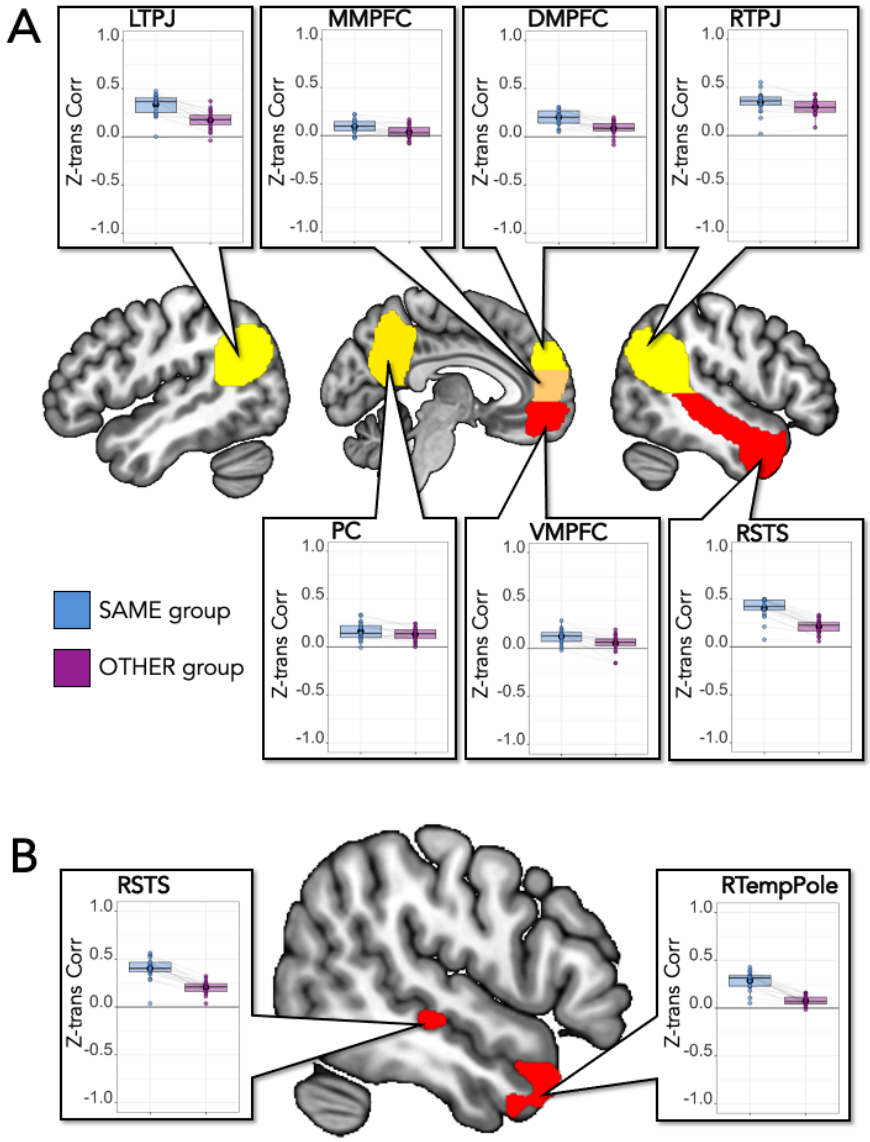


Figure 9: SS-IntDialog correlations within theory of mind regions

(A) Center*: Theory of mind parcels overlaid on template brain (parcels include LTPJ, MMPFC, DMPFC, RTPJ, PC, VMPFC, and RSTS from* (Dufour et al., 2013)*). **Panels**: Average z-transformed Pearson's correlation between each held-out subject's timecourse within each ss-*

fROI and the average timecourse of the remaining participants who viewed and listened to the same version of the stimuli (blue) and the average of the participants who heard the opposite audio stream (purple), averaged across two runs. Each individual's datapoints are connected by light gray lines. For all regions except MMPFC and PC, the within-group correlations were higher than the between-group correlations.

***(B) Center**: Right superior temporal sulcus and right temporal pole were activated for [Forward Dialogue > Forward Monologue] > [Backward Dialogue > Backward Monologue]. Threshold p<.001, uncorrected (df=19, two-tailed). Nothing survives at TFCE corrected threshold. Shown here are clusters for Right STS and Right Temporal Pole; not shown: Left STS and Left Cerebellum. **Panels**: Average z-transformed Pearson's correlation between each held-out subject's timecourse within each ss-fROI and the average timecourse of the remaining participants who viewed and listened to the same version of the stimuli (blue) and the average of the participants who heard the opposite audio stream (purple), averaged across two runs. Each individual's datapoints are connected by light gray lines.*

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
DMPFC	M(SD) = 0.20(0.08) ; range = 0.05-0.31 <u>One-sample t-test</u>: t-value = 10.90, **p< .001 ***	M(SD) = 0.09(0.07); range = -0.08-0.20 <u>One-sample t-test</u>: t-value = 5.39, **p< .001 ***	t = 4.54, **p< .001***
MMPFC	M(SD) = 0.09(0.07); range = -0.03-0.22 <u>One-sample t-test</u>: t-value = 5.66, **p< .001***	M(SD) = 0.036(0.07); range = -0.08-0.16 <u>One-sample t-test</u>: t-value = 2.28, p-value = 0.03	t = 2.83, p-value = 0.01
VMPFC	M(SD) = 0.13(0.08); range = -	M(SD) = 0.05(0.09); range	t = 3.66, **p-**

	0.02-0.29 One-sample t-test: t-value = 7.47, **p< .001***	= -0.15-0.20 One-sample t-test: t-value = 2.64, p-value = 0.02	**value = 0.002***
LTPJ	M(SD) = 0.33(0.11); range = -0.0007-0.47 One-sample t-test: t-value = 13.04, **p< .001 ***	M(SD) = 0.17(0.10); range = -0.04-0.37 One-sample t-test: t-value = 7.91, **p< .001 ***	t = 8.13, **p< .001 *****
PC	M(SD) = 0.17(0.09); range = -0.005-0.34 One-sample t-test: t-value = 8.09, **p< .001 ***	M(SD) = 0.13(0.07); range = 0.001-0.24 One-sample t-test: t-value = 8.38, **p< .001 ***	t = 2.00, p-value = 0.06
RTPJ	M(SD) = 0.34(0.11); range = 0.02-0.55 One-sample t-test: t-value = 13.54, **p< .001 ***	M(SD) = 0.30(0.08); range = 0.08-0.43 One-sample t-test: t-value = 16.26, **p< .001 ***	t = 3.02, **p-value = 0.007 ***
RSTS	M(SD) = 0.40(0.11); range = 0.08-0.50 One-sample t-test: t-value = 16.78, **p< .001 ***	M(SD) = 0.21(0.07); range = 0.06-0.34 One-sample t-test: t-value = 12.90, **p< .001***	t = 11.57, **p< .001 ***

Table 7: SS-IntDialog timecourse correlations within theory of mind regions

Average z-transformed Pearson's correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected p-values). Paired t-test shows that there were higher within-group than between-group correlations for each ToM region except MMPFC and PC (uncorrected p-values).

** indicates significance level p<.05, Bonferroni corrected for 7 ROIs (p<.0071)*

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
RTempPole	M(SD) = 0.29(0.10); range = 0.05-0.43 One-sample t-test: t-value = 13.34	M(SD) = 0.08(0.05); range = -0.01-0.16 One-sample t-test: t-value = 6.79	t = 10.28
RSTS	M(SD) = 0.40(0.12); range = 0.04-0.56 One-sample t-test: t-value = 15.45	M(SD) = 0.21(0.06); range = 0.04-0.32 One-sample t-test: t-value = 14.21	t = 10.01
LCere	M(SD) = 0.26(0.08); range = 0.12-0.38 One-sample t-test: t-value = 14.56	M(SD) = 0.17(0.08); range = 0.003-0.28 One-sample t-test: t-value = 9.69	t = 4.0738
LSTS	M(SD) = 0.40(0.14); range = 0.01-0.57 One-sample t-test: t-value = 12.95	M(SD) = 0.03(0.07); range = -0.10-0.12 One-sample t-test: t-value = 1.87	t = 10.16

Table 8: SS-IntDialog timecourse correlations within conversation regions

Average z-transformed correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected). Paired t-test shows higher within-group than between-group correlations within each region. No p-values are reported since analyses were exploratory.

Summary

Even when some of the auditory content of dialogue was removed by alternating comprehensible and incomprehensible speech across interleaved speakers, canonical left-hemisphere cortical language regions still responded more to forward than backward speech. Furthermore, the

timecourses of activity in canonical left-hemisphere language regions were more similar among individuals listening to the same speech input, compared to individuals listening to the opposite auditory input, holding constant all visual input from the dialogue videos. Notably, there was little to no similarity in the response of canonical language regions for individuals listening to opposite auditory streams, suggesting that language regions were insensitive to any other aspects of the dialogue stimuli that were the same across participants. On the other hand, multiple theory of mind regions and right hemisphere homologues of language regions were correlated even when participants were listening to the opposite auditory streams, suggesting that these regions tracked similarities in the dialogue videos in addition to just the comprehensibility of the speech (though all except right IFG, right AngG, MMPFC, and PC still had higher correlations when participants were listening to the same speech stream).

General Discussion

In this study, we tested the scope and limits of language regions' function by probing them with naturalistic videos of dialogue speech. Canonical language network is highly specific to language processing (Braga et al., 2020; Fedorenko et al., 2010, 2011; Ivanova et al., 2020; Liu et al., 2020; Pritchett et al., 2018). Dialogue, compared to monologue, is a useful boundary test case: tracking multiple speakers and perspectives is part of comprehending language in dialogue, but it is not a function that has been attributed to language regions.

In two tasks, we manipulated the audio stream of *Sesame Street* videos to create matched segments of videos with forward (comprehensible) and backward (incomprehensible) speech. Based on this manipulation, we defined three measures of a cortical region's (in)sensitivity to the

dialogue context of linguistic input. First, a region that processes language independent of a dialogue context should respond equally robustly to forward speech whether presented as a monologue or dialogue. Second, it should respond selectively to the comprehensible speech segments in a dialogue that alternates between forward and backwards speech, even within the frequent alternations of dialogue that render some utterances quite short. Third, the reliable (between-participants) timecourse of response to these alternating dialogue stimuli should be driven only by the timing of the comprehensible speech segments, and not by any other features of the dialogue. By all three of these measures, we find that left hemisphere canonical language regions are insensitive to whether language is in the form of dialogue during passive observation.

Insensitivity to dialogue in canonical language regions

We chose two different analytic approaches to ascertain whether language regions are sensitive to the back-and-forth speaker alternation in dialogue. Using a blocked design, we found no differences in the magnitude of neural response to dialogue versus monologue in the core left-hemisphere language network, though other cortical regions did show enhanced responses during dialogue (see below). With the longer audiovisual stimuli in the SS-IntDialog task, we employed a different analytic technique (ISC) that is more commonly used for naturalistic fMRI datasets (Hasson et al., 2004). ISC analyses allowed us to ask: does anything other than language comprehensibility, at the level of individual utterances, drive the neural responses in language regions during a dialogue? Participants who watched the exact same video clips had similar responses to the stimuli in language regions (i.e., positive within-group correlations within language regions). Critically, though, participants who watched the same video clips with the opposite characters speaking forward versus backward showed close to zero correlation in

canonical language network activity. Thus, nothing else about the dialogue, other than the comprehensibility of the speech stream, was reliably tracked by canonical language regions across participants. This did not have to be the case, as we saw in other regions (discussed below).

Our two analytic approaches, and two task designs, both produced results that converged on a single conclusion: canonical language regions are not sensitive features of dialogue other than language. This insensitivity is consistent with other evidence that language regions are sensitive to relatively local linguistic features, and with evidence that canonical language regions have fairly short temporal receptive windows (Blank & Fedorenko, 2020).

Sensitivity to dialogue outside language regions

In addition to canonical language regions, we also asked whether other regions might be specifically sensitive to dialogue. First, we looked in specific regions that one might expect to respond differently to dialogue and monologue: ToM regions and right hemisphere homologues of language regions. Some regions in both groups fulfilled our criteria for responding preferentially to dialogue.

Most individually-defined ToM regions responded more during forward than backward speech, and more during two-character dialogue interactions than a single character speaking. Two regions responded most during comprehensible dialogue, as shown by an interaction of language and video type. As converging evidence, in SS-IntDialog, activity in ToM regions was also correlated across participants even when they were listening to the opposite speech streams in

Experiment 2 (with the exception of MMPFC and VMPFC), which suggests that aspects of the stimuli other than comprehensibility influence activity in these regions. In experiments using single source texts, ToM regions respond selectively to stimuli that describe or imply contrasting beliefs, knowledge or emotions, between characters or over time (Dodell-Feder et al., 2011; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). Naturalistic dialogue often implies differences of perspective, as speakers use utterances to show how a prior utterance was or was not understood, and to reveal and correct gaps in common ground. Previous work has shown that activity in ToM regions can be synchronized with language regions during language comprehension (Paunov et al., 2019) even though these networks are functionally distinct (Paunov et al., 2022; Shain et al., 2022); thus, it is unsurprising that ToM network was engaged in processing dialogue in our task.

The other set of regions we examined was right language homologues. We localized these regions in the same way as the left hemisphere regions - using the independent language localizer task, selecting top voxels from parcels in the right hemisphere. All of these regions did respond more to comprehensible than incomprehensible speech, and 2 of the 6 regions also responded more to dialogue than monologue. More intriguingly, all of the right hemisphere language homologues showed significant correlations in SS-IntDialog even between participants listening to opposite versions of the stimuli. Thus, dialogue videos evoke enhanced activity in some right hemisphere, but not left hemisphere, regions defined by their language selectivity. Consistent with these results, previous work has demonstrated that pragmatic and social aspects of language may be processed by regions in the right hemisphere. For instance, processing emotional prosody has been shown to be right lateralized (Friederici, 2011; Frühholz et al., 2012; Ross & Monnot, 2008; Seydell-Greenwald et al., 2020), and regions responsive to prosody

differences have been shown to be distinct from language regions, even among individuals with large perinatal strokes in the left hemisphere whose language regions are located in the right hemisphere (Newport et al., 2022). Right hemisphere damage can make it more difficult for individuals to make inferences from discourse (Beeman, 1993).

Finally, in addition to looking within specific regions, we also examined responses across the whole brain to determine where comprehensible dialogue specifically led to higher activation. Significant clusters were identified in right temporal pole, right STS, left STS, and left cerebellum (though note that none of these survived correction for multiple comparisons). While these results are exploratory, these regions may be useful targets for future studies on dialogue comprehension. Part of right temporal pole, for example, has long been thought to be involved in social and emotional processing, among other higher level cognitive functions (Herlin et al., 2021; Olson et al., 2007; Pehrs et al., 2017; Wakusawa et al., 2007), and parts of the cerebellum are also involved in language and social cognition (D'Mello & Stoodley, 2015; Stoodley, 2012; Van Overwalle et al., 2014); thus, it is plausible that the clusters we identified in these regions might be meaningful subregions involved in dialogue comprehension. Right STS in particular is a key region supporting social interaction processing, though it is important to note that this is a large region with multiple subregions subserving different functions (Deen et al., 2015). Parts of STS respond to visual social interactions (Walbrin et al., 2018; Walbrin & Koldewyn, 2019), and in particular, a specific part of posterior STS responds to interactions between agents (Isik et al., 2017). Using a naturalistic dataset, part of STS was shown to be selective for interaction, separate from TOM (Lee Masson & Isik, 2021). Other evidence points to additional roles of STS regions in social processing, such as directing attention (Materna et al., 2008) and processing

prosody (Wildgruber et al., 2006). Given the interactive nature of the dialogue stimuli, it makes sense that part of STS might be involved in processing the social interaction that occurs in a comprehensible dialogue, either as a subregion that responds to both social interaction and voices, or because the content of the language enhances the perception of an interaction.

Limitations

In these experiments, the stimuli were experimentally manipulated clips from professionally produced episodes of the television show, *Sesame Street*. This stimulus source had both strengths and limitations, in terms of experimental design for testing the function of language regions. First, there were some differences in the linguistic features between the dialogue and monologue clips (**Supplementary Figure 4**). For example, monologues included longer sentences and more complex speech than dialogue. Second, we used puppets with rigid mouths, to allow us to align backwards speech with the video stimuli. Both children and adults are highly sensitive to cross modal misalignment of lip movements and speech sounds. However, there may be residual differences between conditions in the audiovisual alignment, since the puppets were originally filmed to match the forward speech stream. Third, we used backwards speech as the control condition rather than acoustically degraded speech (Overath et al., 2015; Stoppelman et al., 2013) or foreign speech (Schlosser et al., 1998). We chose to use backwards speech because it (1) allowed us to maintain the continuity of voice within character, (2) it sounded more natural than the degraded speech when embedded in the video, and (3) most critically, because it allowed us to control for the visual stimuli across participants (e.g., we could have different participants watch the exact same clips, either forward or backward), which would not have been feasible given the availability of non-English versions of the selected clips. Future studies could

design stimuli with comprehensible and foreign speech using bilingual actors to match the visual input and voices.

Future Directions

A clear extension of this work - and indeed, the motivation behind it - is to use these language tasks with young children. Extensive prior literature has demonstrated the benefits of naturalistic movie-based stimuli for young children (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2015, 2019). It was non-trivial that this quasi-naturalistic approach worked at all for univariate analyses: one concern we had was that the uncontrolled elements of the video clips would prevent us from extracting a robust language response. This was not the case: not only did our block design in Experiment 1 elicit a response with substantial overlap with a more traditional language localizer - both at the group and individual level - but even the naturalistic alternating dialogue in Experiment 2 reliably localized language regions. In each participant, the same voxels were identified by the forward vs backward manipulation of the *Sesame Street* clips as by a well-validated standard auditory language localizer task (Scott et al., 2017). Furthermore, this child-friendly task was engaging and effective for adults, suggesting that it holds promise as a task that can be used across a wide age range. Thus, this task may be useful for other populations that may find classic language tasks hard to tolerate, such as individuals with developmental or acquired disorders.

Conclusions

Our results suggest that canonical left hemisphere language regions are not sensitive to aspects of dialogue interactions other than comprehensibility of the speech stream, even though other

aspects of dialogue can impact language comprehension. Furthermore, we found that embedding an explicit experimental control for language (in our case, backwards speech) within a naturalistic context (in our case, clips from *Sesame Street*) is a feasible, engaging, and robust approach for studying language processing in the brain.

Materials

Stimuli transcriptions and descriptions, analysis code, stimulus presentation code, processed data, and link to raw data on OpenNeuro can be found on OSF: https://osf.io/whsb7/. Raw stimuli can be provided upon request.

References

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the Language Given: The Neural Correlates of Inferring Speaker Meaning. *Cerebral Cortex*, *24*(10), 2572–2578. https://doi.org/10.1093/cercor/bht112

Bates, E., Reilly, J., Wulfeck, B., Dronkers, N., Opie, M., Fenson, J., Kriz, S., Jeffries, R., Miller, L., & Herbst, K. (2001). Differential Effects of Unilateral Lesions on Language Production in Children and Adults. *Brain and Language*, *79*(2), 223–265. https://doi.org/10.1006/brln.2001.2482

Beeman, M. (1993). Semantic Processing in the Right Hemisphere May Contribute to Drawing Inferences from Discourse. *Brain and Language*, *44*(1), 80–120. https://doi.org/10.1006/brln.1993.1006

Bendtz, K., Ericsson, S., Schneider, J., Borg, J., Bašnáková, J., & Uddén, J. (2022). Individual Differences in Indirect Speech Act Processing Found Outside the Language Network. *Neurobiology of Language*, *3*(2), 287–317. https://doi.org/10.1162/nol_a_00066

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *17*(1), 353–362. https://doi.org/10.1523/JNEUROSCI.17-01-00353.1997

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925. https://doi.org/10.1016/j.neuroimage.2020.116925

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

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013

Blumstein, S. E., & Amso, D. (2013). Dynamic Functional Organization of Language: Insights From Functional Neuroimaging. *Perspectives on Psychological Science*, *8*(1), 44–48. https://doi.org/10.1177/1745691612469021

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

Broca, P. (1865). Sur le siège de la faculté du langage articulé. *Bulletins et Mémoires de la Société d'Anthropologie de Paris*, *6*(1), 377–393. https://doi.org/10.3406/bmsap.1865.9495

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, *216*, 116464. https://doi.org/10.1016/j.neuroimage.2019.116464

Cantlon, J. F., & Li, R. (2013). Neural Activity during Natural Viewing of Sesame Street Statistically Predicts Test Scores in Early Childhood. *PLOS Biology*, *11*(1), e1001462. https://doi.org/10.1371/journal.pbio.1001462

Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language*, *92*, 234–253. https://doi.org/10.1016/j.jml.2016.06.013

Chang, L., Manning, J., Baldassano, C., Vega, A. de la, Fleetwood, G., Geerligs, L., Haxby, J., Lahnakoski, J., Parkinson, C., Shappell, H., Shim, W. M., Wager, T., Yarkoni, T., Yeshurun, Y., & Finn, E. (2020, July 9). *naturalistic-data-analysis/naturalistic_data_analysis: Version 1.0*. https://doi.org/10.5281/zenodo.3937849

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*(2), 259–294. https://doi.org/10.1016/0364-0213(89)90008-6

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

D'Mello, A. M., & Stoodley, C. J. (2015). Cerebro-cerebellar circuits in autism spectrum disorder. *Frontiers in Neuroscience*, *9*. https://www.frontiersin.org/articles/10.3389/fnins.2015.00408

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). FMRI item analysis in a theory of mind task. *NeuroImage*, *55*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1), 145–177. https://doi.org/10.1016/j.cognition.2003.11.002

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D. E., & Saxe, R. (2013). Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE*, *8*(9), e75468. https://doi.org/10.1371/journal.pone.0075468

Enge, A., Friederici, A. D., & Skeide, M. A. (2020). A meta-analysis of fMRI studies of language comprehension in children. *NeuroImage*, *215*, 116858. https://doi.org/10.1016/j.neuroimage.2020.116858

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), Article 1. https://doi.org/10.1038/s41592-018-0235-4

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433. https://doi.org/10.1073/pnas.1112937108

Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. *Current Biology*, *22*(21), 2059–2062. https://doi.org/10.1016/j.cub.2012.09.011

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Ivanova, A., Dhamala, R., & Bers, M. U. (2019). The Language of Programming: A Cognitive Perspective. *Trends in Cognitive Sciences*, *23*(7), 525–528. https://doi.org/10.1016/j.tics.2019.04.010

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126. https://doi.org/10.1016/j.tics.2013.12.006

Feng, W., Wu, Y., Jan, C., Yu, H., Jiang, X., & Zhou, X. (2017). Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain and Language*, *171*, 52–61. https://doi.org/10.1016/j.bandl.2017.04.005

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, *27*(1), 35–53. https://doi.org/10.1080/01638539909545049

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Frühholz, S., Ceravolo, L., & Grandjean, D. (2012). Specific Brain Networks during Explicit and Implicit Decoding of Emotional Prosody. *Cerebral Cortex*, *22*(5), 1107–1117. https://doi.org/10.1093/cercor/bhr184

Fusaroli, R., & Tylén, K. (2016). Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. *Cognitive Science*, *40*(1), 145–171. https://doi.org/10.1111/cogs.12251

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old Infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gold, B. T., & Buckner, R. L. (2002). Common Prefrontal Regions Coactivate with Dissociable Posterior Regions during Controlled Semantic and Phonological Tasks. *Neuron*, *35*(4), 803–812. https://doi.org/10.1016/S0896-6273(02)00800-0

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://www.frontiersin.org/articles/10.3389/fninf.2011.00013

Grall, C., & Finn, E. S. (2022). Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience*, *17*(6), 598–608. https://doi.org/10.1093/scan/nsac019

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current Opinion in Neurobiology*, *28*, 136–141. https://doi.org/10.1016/j.conb.2014.07.013

Hagoort, P., & Indefrey, P. (2014). The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, *37*(1), 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Hasson, U., & Honey, C. J. (2012). Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, *62*(2), 1272–1278. https://doi.org/10.1016/j.neuroimage.2012.02.004

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Herlin, B., Navarro, V., & Dupont, S. (2021). The temporal pole: From anatomy to function—A literature appraisal. *Journal of Chemical Neuroanatomy*, *113*, 101925. https://doi.org/10.1016/j.jchemneu.2021.101925

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., Ryan, E., Nieto-Castañón, A., Ferreira, V., & Fedorenko, E. (2022). Precision fMRI reveals that the

language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, bhac350. https://doi.org/10.1093/cercor/bhac350

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. https://doi.org/10.1073/pnas.1714471114

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U.-M., Bers, M. U., & Fedorenko, E. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *ELife*, *9*, e58906. https://doi.org/10.7554/eLife.58906

Jang, G., Yoon, S., Lee, S.-E., Park, H., Kim, J., Ko, J. H., & Park, H.-J. (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage*, *81*, 61–72. https://doi.org/10.1016/j.neuroimage.2013.05.027

Kamps, F. S., Richardson, H., Murty, N. A. R., Kanwisher, N., & Saxe, R. (2022). Using child-friendly movie stimuli to study the development of face, place, and object regions from age 3 to 12 years. *Human Brain Mapping*, *43*(9), 2782–2800. https://doi.org/10.1002/hbm.25815

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

Levinson, S. C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Lewkowicz, D. J., & Flom, R. (2014). The Audiovisual Temporal Binding Window Narrows in Early Childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Liu, Y.-F., Kim, J., Wilson, C., & Bedny, M. (2020). Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *ELife*, *9*, e59340. https://doi.org/10.7554/eLife.59340

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: The neurobiology of sign language. *Trends in Cognitive Sciences*, *12*(11), 432–440. https://doi.org/10.1016/j.tics.2008.07.010

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), Article 8. https://doi.org/10.1038/s41593-022-01114-5

Materna, S., Dicke, P. W., & Thier, P. (2008). The posterior superior temporal sulcus is involved in social communication not specific for the eyes. *Neuropsychologia*, *46*(11), 2759–2765. https://doi.org/10.1016/j.neuropsychologia.2008.05.016

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared Language: Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. *Psychological Science*, *23*(8), 914–922. https://doi.org/10.1177/0956797612437427

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

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), Article 1. https://doi.org/10.1038/s41597-021-01033-3

Neville, H. J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., & Turner, R. (1998). Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of the National Academy of Sciences*, *95*(3), 922–929. https://doi.org/10.1073/pnas.95.3.922

Newport, E. L., Seydell-Greenwald, A., Landau, B., Turkeltaub, P. E., Chambers, C. E., Martin, K. C., Rennert, R., Giannetti, M., Dromerick, A. W., Ichord, R. N., Carpenter, J. L., Berl, M. M., & Gaillard, W. D. (2022). Language and developmental plasticity after perinatal stroke. *Proceedings of the National Academy of Sciences*, *119*(42), e2207293119. https://doi.org/10.1073/pnas.2207293119

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, *130*(7), 1718–1731. https://doi.org/10.1093/brain/awm052

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), Article 6. https://doi.org/10.1038/nn.4021

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (2022). Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00071

Pehrs, C., Zaki, J., Schlochtermeier, L. H., Jacobs, A. M., Kuchinke, L., & Koelsch, S. (2017). The Temporal Pole Top-Down Modulates the Ventral Visual Stream During Social Cognition. *Cerebral Cortex*, *27*(1), 777–792. https://doi.org/10.1093/cercor/bhv226

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*(1), 62–88. https://doi.org/10.1111/j.1749-6632.2010.05444.x

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

Pritchett, B. L., Hoeflin, C., Koldewyn, K., Dechter, E., & Fedorenko, E. (2018). High-level language processing regions are not engaged in action observation or imitation. *Journal of Neurophysiology*, *120*(5), 2555–2570. https://doi.org/10.1152/jn.00222.2018

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03399-2

Ross, E. D., & Monnot, M. (2008). Neurology of affective prosody and its functional–anatomic organization in right hemisphere. *Brain and Language*, *104*(1), 51–74. https://doi.org/10.1016/j.bandl.2007.04.007

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382. https://doi.org/10.1353/lan.1977.0041

Schlosser, M. J., Aoyagi, N., Fulbright, R. K., Gore, J. C., & McCarthy, G. (1998). Functional MRI studies of auditory comprehension. *Human Brain Mapping*, *6*(1), 1–13. https://doi.org/10.1002/(SICI)1097-0193(1998)6:1<1::AID-HBM1>3.0.CO;2-7

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Seydell-Greenwald, A., Chambers, C. E., Ferrara, K., & Newport, E. L. (2020). What you say versus how you say it: Comparing sentence comprehension and emotional prosody processing using fMRI. *NeuroImage*, *209*, 116509. https://doi.org/10.1016/j.neuroimage.2019.116509

Shain, C., Paunov, A., Chen, X., Lipkin, B., & Fedorenko, E. (2022). *No evidence of theory of mind reasoning in the human language network* (p. 2022.07.18.500516). bioRxiv. https://doi.org/10.1101/2022.07.18.500516

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Stoodley, C. J. (2012). The Cerebellum and Cognition: Evidence from Functional Imaging Studies. *The Cerebellum*, *11*(2), 352–365. https://doi.org/10.1007/s12311-011-0260-7

Stoppelman, N., Harpaz, T., & Ben-Shachar, M. (2013). Do not throw out the baby with the bath water: Choosing an effective baseline for a functional localizer of speech processing. *Brain and Behavior*, *3*(3), 211–222. https://doi.org/10.1002/brb3.129

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797. https://doi.org/10.1073/pnas.94.26.14792

Tolins, J., & Fox Tree, J. E. (2016). Overhearers Use Addressee Backchannels in Dialog Comprehension. *Cognitive Science*, *40*(6), 1412–1434. https://doi.org/10.1111/cogs.12278

Van Overwalle, F., Baetens, K., Mariën, P., & Vandekerckhove, M. (2014). Social cognition and the cerebellum: A meta-analysis of over 350 fMRI studies. *NeuroImage*, *86*, 554–572. https://doi.org/10.1016/j.neuroimage.2013.09.033

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Wakusawa, K., Sugiura, M., Sassa, Y., Jeong, H., Horie, K., Sato, S., Yokoyama, H., Tsuchiya, S., Inuma, K., & Kawashima, R. (2007). Comprehension of implicit meanings in social situations involving irony: A functional MRI study. *NeuroImage*, *37*(4), 1417–1426. https://doi.org/10.1016/j.neuroimage.2007.06.013

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*, 31–39. https://doi.org/10.1016/j.neuropsychologia.2018.02.023

Walbrin, J., & Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. *NeuroImage*, *198*, 296–302. https://doi.org/10.1016/j.neuroimage.2019.05.027

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty

Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Wildgruber, D., Ackermann, H., Kreifelts, B., & Ethofer, T. (2006). Cerebral processing of linguistic and emotional prosody: FMRI studies. In S. Anders, G. Ende, M. Junghofer, J. Kissler, & D. Wildgruber (Eds.), *Progress in Brain Research* (Vol. 156, pp. 249–268). Elsevier. https://doi.org/10.1016/S0079-6123(06)56013-3

Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., & Eriksson, D. K. (2017). Validity and reliability of four language mapping paradigms. *NeuroImage: Clinical*, *16*, 399–408. https://doi.org/10.1016/j.nicl.2016.03.015