OXFORD

Cerebral Cortex, 2022, 1–21
https://doi.org/10.1093/cercor/bhac505
Original Article

No evidence of theory of mind reasoning in the human language network

Cory Shain[1,*,†], Alexander Paunov[2,†], Xuanyi Chen[3,†], Benjamin Lipkin[1], Evelina Fedorenko[1,4]

[1]Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, MIT Bldg 46-316077 Massachusetts Avenue, Cambridge, MA 02139, United States,
[2]INSERM-CEA Cognitive Neuroimaging Unit (UNICOG), NeuroSpin Center, Gif sur Yvette 91191, France,
[3]Department of Cognitive Sciences, Rice University, 6100 Main Street, Houston, TX 77005, United States,
[4]Program in Speech Hearing in Bioscience and Technology, Harvard Medical School, 260 Longwood Avenue, TMEC 333, Boston, MA 02115, United States
*Corresponding author: Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, MIT, Cambridge, MA 46-4127, United States. Email: cshain@mit.edu
†Cory Shain, Alexander Paunov and Xuanyi Chen contributed equally.

Language comprehension and the ability to infer others' thoughts (theory of mind [ToM]) are interrelated during development and language use. However, neural evidence that bears on the relationship between language and ToM mechanisms is mixed. Although robust dissociations have been reported in brain disorders, brain activations for contrasts that target language and ToM bear similarities, and some have reported overlap. We take another look at the language-ToM relationship by evaluating the response of the language network, as measured with fMRI, to verbal and nonverbal ToM across 151 participants. Individual-participant analyses reveal that all core language regions respond more strongly when participants read vignettes about false beliefs compared to the control vignettes. However, we show that these differences are largely due to linguistic confounds, and no such effects appear in a nonverbal ToM task. These results argue against cognitive and neural overlap between language processing and ToM. In exploratory analyses, we find responses to social processing in the "periphery" of the language network—right-hemisphere homotopes of core language areas and areas in bilateral angular gyri—but these responses are not selectively ToM-related and may reflect general visual semantic processing.

Key words: fMRI; language; theory of mind; right hemisphere; angular gyrus.

Introduction

Everyday social interactions regularly involve an intricate coordination between language use on the one hand and reasoning about others' mental states (theory of mind [ToM]) on the other. For example, to understand the communicative intent behind an utterance (e.g. "Nice outfit"), we often rely on inferences about the speaker's state of mind (e.g. whether they are likely to have a positive assessment of your outfit and thus whether the utterance was likely sincere or sarcastic). In addition to this kind of ToM-based pragmatic reasoning needed to infer implicit meanings from utterances (e.g. Grice 1975; Sperber and Wilson 1987; Winner et al. 1998; Champagne-Lavau and Joanette 2009; Roberts 2012), linguistic representations may be critical to the development of ToM (e.g. Astington and Jenkins 1999; Peterson and Siegal 2000; Hale and Tager-Flusberg 2003; Ruffman et al. 2003; Astington and Baird 2005; Slade and Ruffman 2005; Miller 2006; de Villiers and de Villiers 2014; Richardson et al. 2020). There is thus reason to suspect a close cognitive and neural connection between language and ToM processing and perhaps even overlap between the neural resources that support both kinds of skills.

However, neuroscientific evidence that bears on the relationship between language and ToM paints a complex picture. On the one hand, at least some evidence indicates that language and ToM rely on distinct cognitive and neural mechanisms. In particular, ToM reasoning abilities can be preserved in cases of linguistic deficits (e.g. in aphasia; e.g. Dronkers et al. 1998; Varley et al. 2001; Apperly et al. 2006; Willems et al. 2011), and at least some aspects of language can be preserved when social abilities are impaired (e.g. in some individuals with autism spectrum disorders; e.g. Tager-Flusberg et al. 2005; Diehl et al. 2006). Furthermore, the core brain areas that have been linked to language vs. ToM appear to be distinct. Language processing recruits a left-lateralized network of lateral frontal and temporal areas (e.g. Binder et al. 1997; Fedorenko et al. 2010), whereas social cognitive processing, including ToM/mentalizing, recruits bilateral areas (though more strongly present in the right hemisphere [RH]) at the junction of temporal and parietal cortexes along with frontal and parietal cortical midline regions (e.g. Fletcher et al. 1995; Castelli et al. 2000; Gallagher et al. 2000; Vogeley et al. 2001; Ruby and Decety 2003; Saxe and Kanwisher 2003, inter alia). These sets of areas also dissociate during naturalistic cognition: They show strong within-network correlations and weaker correlations among pairs of brain regions that straddle network boundaries (e.g. Paunov et al. 2019; Braga et al. 2020) and "track" different aspects of naturalistic stimuli (Paunov et al. 2022). On the other hand, whole-brain activation landscapes for contrasts that target language processing and those that target ToM bear similarities (Fig. 1). Further, Deen et al. (2015; cf. Koster-Hale and Saxe 2013) examined responses to language and ToM using individual-participant analyses and reported partial overlap between language and ToM areas in the posterior temporal lobe and angular gyrus. But Deen et al.'s study used a ToM contrast based on verbal vignettes that could have linguistic differences, making these findings difficult to interpret.

In an effort to clarify the relationship between the language and the ToM networks, we examine responses in the frontal and temporal language areas to the standard verbal ToM contrast (false belief stories > false photograph stories; Saxe and Kanwisher 2003; the same contrast as was used in

Received: July 18, 2022. **Revised:** November 30, 2022. **Accepted:** December 1, 2022

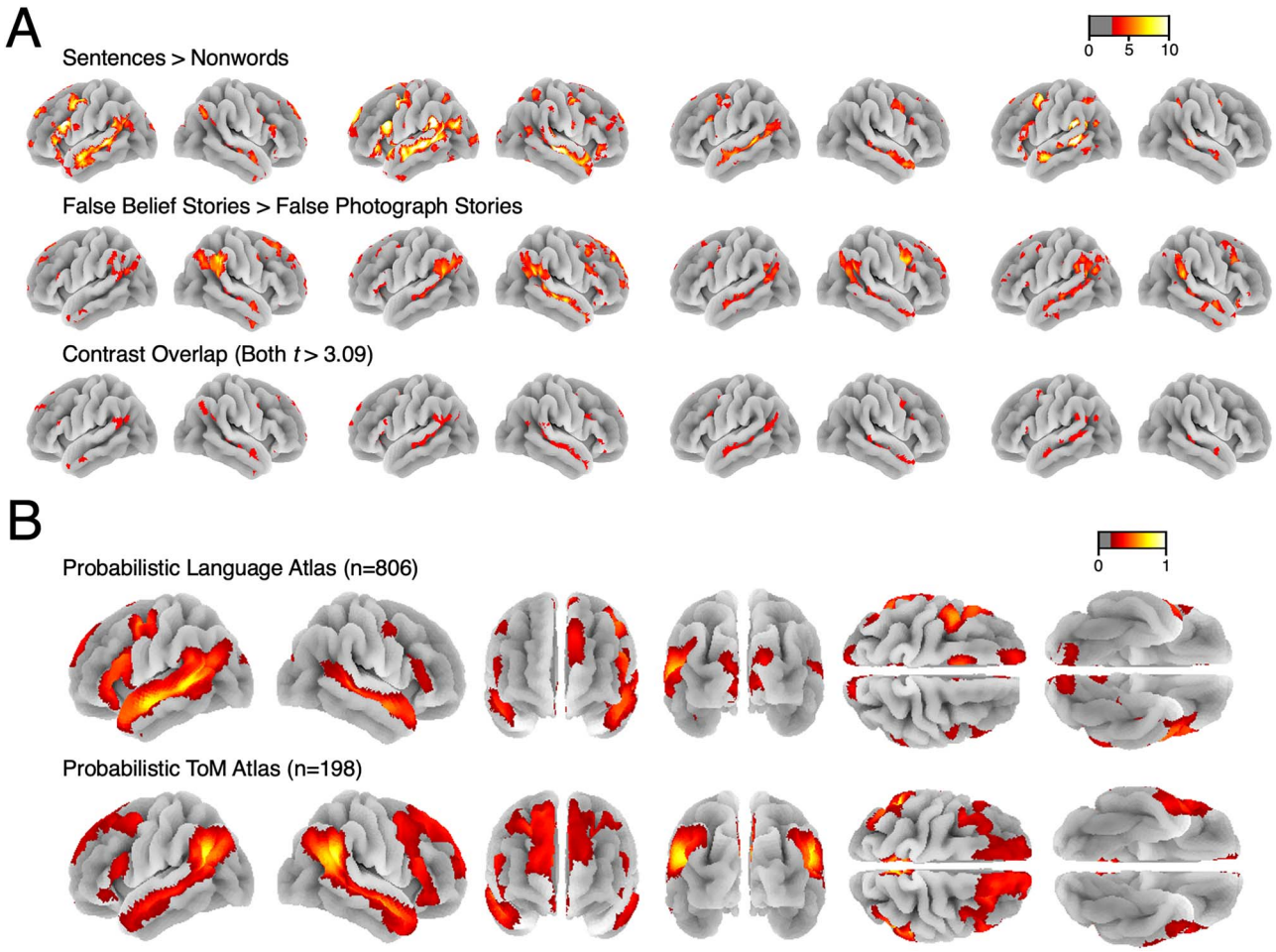

Fig. 1. Comparison of whole-brain activation patterns for language (sentences > nonwords) and ToM (false belief > false photo) contrasts. A) Responses to the language (top row) and ToM (middle row) localizer contrasts in 4 sample participants. Overlap (bottom row) is observed primarily in and around the angular gyrus/TPJ and along the STS, with some scattered overlap in the lateral frontal cortex. B) Whole-brain probabilistic atlases for the language and ToM localizer contrasts created from 2 large fMRI datasets by overlaying individual activation maps (see Lipkin et al. 2022 for details). The 2 tasks elicit broadly similar spatial distributions of activity.

Deen et al. 2015) but also to a nonverbal ToM contrast (mentalizing > nonmentalizing events in a rich naturalistic stimulus—a few-minute-long Pixar film; Jacoby et al. 2016). Jacoby et al. (2016) have previously shown that this nonverbal ToM localizer elicits a strong response in brain areas defined by the verbal ToM localizer (see also Richardson et al. 2018; Kamps et al. 2022).

In addition to our critical question about the involvement of core language network areas in ToM processing, we conduct exploratory analyses to investigate possible ToM responses in brain areas in the "periphery" of the language network—areas that show some response during language processing but are less strongly integrated with the core left-hemisphere (LH) language areas than those core areas are with one another (Fedorenko and Thompson-Schill 2014; Chai et al. 2016; Fig. 2). These peripheral regions include the RH homotopes of the frontal and temporal LH areas and bilateral areas in the angular gyrus and have been shown to differ in their functional profiles from the core language areas (e.g. Blank, Balewski, et al. 2016a; Blank, Duff, et al. 2016b; Fedorenko et al. 2020; Ivanova et al. 2020, inter alia) and to show reduced static and dynamic functional correlations with the core language areas (e.g. Blank et al. 2014; Chai et al. 2016; Paunov et al. 2019; Braga et al. 2020). Further, these broad anatomical areas have been implicated by prior work in ToM, social processing, or social/affective aspects of language processing: the RH lateral frontal and temporal areas (e.g. Kaplan et al. 1990; Winner et al. 1998; Mitchell and Crow 2005; Rajimehr et al. 2022; Hauptman et al. 2022), as well as bilateral angular gyri (e.g. Saxe and Kanwisher 2003; Saxe 2006, 2010; Lombardo et al. 2011; Mar 2011; Schurz et al. 2014, 2017). These peripheral language areas may therefore show greater functional overlap with ToM reasoning and thus possibly serve as transitional zones between the language-selective and the ToM-selective networks.

To foreshadow our results, we do not find that the core LH language areas support ToM reasoning. Although, similar to Deen et al. (2015), we find that the language areas respond to the verbal ToM contrast, we show (i) that this effect is at least in part due to linguistic differences between the 2 conditions and (ii) that a nonverbal ToM condition does not engage the language network. In the language periphery, we find that nonverbal ToM elicits a strong response in both the RH homotopes of the language areas and in bilateral language-responsive areas in the angular gyrus. However, the detailed response profile of these areas differs from that of the ToM areas. Unlike the ToM areas, these peripheral language areas respond at least as strongly to depictions of social interactions with no mental state content as they do to depictions that encourage mentalizing. These results are therefore

consistent with a broadly social function for peripheral language areas but do not support a role in ToM specifically.

Materials and methods

General approach

Our research design is informed by extensive prior evidence that the language and ToM networks are dissociable functional units in the human brain. First, a range of materials, tasks, and presentation formats yield remarkably stable definitions of both the language network (e.g. Fedorenko et al. 2010; Scott et al. 2017; Malik-Moraleda et al. 2022) and the ToM network (Fletcher et al. 1995; Gallagher et al. 2000; Castelli et al. 2002; Sommer et al. 2007; Mason and Just 2011; Jacoby et al. 2016, see e.g. Koster-Hale and Saxe 2013, for review). Further, these networks emerge from task-free (resting-state) functional correlation data (Braga and Buckner 2017; Braga et al. 2020; DiNicola et al. 2020). These 2 networks generally show little spatial overlap, with areas of the superior temporal sulcus (STS) being the principal exception (Deen et al. 2015; Paunov et al. 2019). Second, the language and ToM networks show high within-network synchrony and lower between-network synchrony during both resting-state and naturalistic language comprehension tasks (Paunov et al. 2019, 2022; Braga et al. 2020), supporting a functional dissociation between them. Third, language and ToM abilities show dissociable patterns of impairment: damage to the language network can impair language processing without impairing ToM reasoning (e.g. Dronkers et al. 1998; Varley et al. 2001; Apperly et al. 2006; Willems et al. 2011), whereas damage to the ToM network can impair ToM reasoning without impairing language processing (e.g. Apperly et al. 2004; Martín-Rodríguez and León-Carrión 2010; Domínguez et al. 2019), and language can be preserved in individuals whose ToM reasoning is otherwise impaired, as in some cases of autism spectrum disorders (e.g. Tager-Flusberg et al. 2005; Diehl et al. 2006) or schizophrenia (Sprong et al. 2007).

Based on the foregoing evidence, in this work, we assume the existence and (at least partial) functional dissociation of the language and ToM networks. We simply use "localizer task" contrasts (described below) as an efficient method to identify these networks in individual brains in order to ask whether effects of interest are present (albeit to a lesser extent) in each network (e.g. whether the language network shows evidence of mentalizing).

Experimental design

Participants

162 individuals from the Cambridge/Boston, MA community participated for payment. All participants completed the language localizer (Fedorenko et al. 2010). They also all completed a verbal ToM localizer task (Saxe and Kanwisher 2003; task details are given in Materials and procedure section); 160 of the 162 participants completed 2 runs of the verbal ToM task, and the remaining 2 participants completed a single run of the verbal ToM task. For the 160 participants who performed 2 runs, we evaluated the quality of the verbal ToM task data by examining the stability of the activation landscape across runs. In particular, we computed an across-runs spatial correlation (Pearson correlation between the voxelwise localizer contrast estimates in each run). This analysis was performed across voxels that fall within the set of ToM masks corresponding to broad areas within which most participants show ToM responses (as described in Materials and procedure) and the correlation values were averaged across the ToM masks to derive a single value per participant. Based on this analysis, we excluded 11 participants with negative spatial correlation values (which suggest poor data quality), leaving 149 participants. For the 2 participants who performed a single run of the verbal ToM task, we evaluated data quality by visual examination of the whole-brain activation maps for the localizer contrast (i.e. false belief > false photo; task details below); both participants' maps looked as expected. Thus, overall, we include 151 participants in the analyses reported here (age = 18–48, mean = 24.7; 99 [66%] female). A subset of these participants ($n = 48$) additionally completed a nonverbal ToM localizer task (Jacoby et al. 2016; task details below). Of these, 34 participants completed both ToM localizer tasks within the same scanning session, whereas the remaining 14 participants completed them in different sessions. Because at least 2 runs of a task are necessary to estimate the response magnitudes to the conditions of that task (to ensure independence between the data used to define the regions of interest and the data used to estimate the responses, as described in Materials and procedure section), the 2 participants with a single run of the verbal ToM task were not included in the analyses of the verbal ToM task, but they could still be used for defining the ToM fROIs and examining the responses in those fROIs to the conditions of the nonverbal ToM task.

138 of the 151 participants (~91%) were right-handed, as determined by the Edinburgh handedness inventory (Oldfield 1971), or self-report; the remaining participants (10 left-handed and 3 ambidextrous) showed typical left-lateralized language activations in the language localizer task (see Willems et al. 2014 for arguments for including left-handers in cognitive neuroscience experiments); 135 participants (~89%) were native English speakers, and the remaining 16 were fluent in English (see Malik–Moraleda et al. 2022 for evidence that language responses are similar in native and fluent speakers of English). All participants gave written informed consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects.

Materials and procedure

For all participants, the scanning sessions lasted approximately 2 h and included some tasks not related to the results reported here.

Language localizer

This task is described in detail in Fedorenko et al. (2010) and subsequent studies from the Fedorenko lab (e.g. Fedorenko et al. 2011, 2020; Blank et al. 2014; Blank, Balewski, et al. 2016a; Blank, Duff, et al. 2016b; Pritchett et al. 2018; Paunov et al. 2019; Shain et al. 2020, among others; available for download from https://evlab.mit.edu/funcloc/). The language localizer targets higher-level aspects of language, including lexical and phrasal semantics, morphosyntax, and sentence-level pragmatic processing, to the exclusion of perceptual (speech- or reading-related) and articulatory processes (see Fedorenko and Thompson-Schill 2014 for discussion). Briefly, participants read sentences and lists of unconnected, pronounceable nonwords in a blocked design with a counterbalanced order across runs. Stimuli were presented 1 word/nonword at a time at the rate of 450 ms per word/nonword. Participants read the materials passively and performed a simple button-press task at the end of each trial (included in order to help participants remain alert). Each block consisted of 3 6-s trials for a total block duration of 18 s. Each run consisted of 8 blocks per condition, with 5 14-s rest periods (1 at the beginning of the run, 1 at the end, and 3 interleaved between blocks). This localizer task has been extensively validated and is shown to be robust to variation in the materials, modality of presentation, language, and task

(Fedorenko et al. 2010; Scott et al. 2017; Malik-Moraleda et al. 2022; Ivanova et al. in prep.). Each participant completed 2 5 m 58 s runs.

ToM localizer (verbal)

This task is described in detail in Saxe and Kanwisher (2003) and in subsequent studies from the Saxe lab (e.g. Saxe and Wexler 2005; Young et al. 2010; Bruneau, Pluta, et al. 2012b; among others; available for download from http://saxelab.mit.edu/use-our-efficient-false-belief-localizer). The verbal ToM localizer targets "representational ToM" (Saxe 2006), akin to "cognitive ToM" (Shamay-Tsoory et al. 2010; Dennis et al. 2013), that is, inferences about the propositional content of agents' beliefs, desires, etc., to the exclusion of "affective ToM," roughly, the capacity to understand and empathize with others' emotional states (e.g. Brothers and Ring 1992; Hein and Singer 2008; Singer and Lamm 2009). The task is based on the classic false belief paradigm (Wimmer and Perner 1983; Wellman et al. 2001) and contrasts verbal vignettes about false beliefs (e.g. a protagonist has a false belief about an object's location; the critical condition) vs. linguistically similar vignettes about false photo states (physical representations depicting outdated scenes, e.g. a photograph showing an object that has since been removed; the control condition). Participants read these vignettes, one at a time, in a slow event-related design. Each vignette was followed by a true/false comprehension question. This localizer task has been extensively validated and has been shown to be robust to the variation in the materials, modality of presentation, and task (Saxe and Kanwisher 2003; Saxe and Wexler 2005; Saxe and Powell 2006; Saxe et al. 2006; Young et al. 2010; Dodell-Feder et al. 2011; Bruneau, Dufour, et al. 2012a; Koster-Hale and Saxe 2013); 149 participants completed 2 runs, each lasting for 4 m 22 s and consisting of 5 vignettes per condition. The remaining 2 participants who completed 1 run were excluded from the analysis of verbal ToM localizer activations but were used for the analysis of responses of the ToM areas to the conditions of the nonverbal ToM localizer.

ToM localizer (nonverbal)

This nonverbal paradigm, based on a silent animated film, is described in detail in Jacoby et al. (2016) and in subsequent studies (e.g. Richardson et al. 2018, 2020; Paunov et al. 2019, 2022; Kamps et al. 2022). Similarly to the verbal ToM localizer, it targets brain regions that support inferences about others' mental states, but in contrast to the main localizer, it is nonverbal, relying on participants engaging in mental state attribution from observed intentional actions. The task consists of passive viewing of an animated short film, Partly Cloudy (Pixar Animation Studios), which contains (i) sections that are likely to elicit mental state attribution—the "mental" condition (e.g. a character falsely believes they have been abandoned by a companion; 4 events, 44 s total); (ii) sections that simply depict physical events—the "physical" condition (e.g. a flock of storks flying; 3 events, 24 s total); (iii) sections that depict characters interacting without strong mental or emotional dimensions—the "social" condition (e.g. a cloud and a stork playing; 5 events, 28 s total); and (iv) sections that depict characters experiencing physical pain—the "pain" condition (e.g. a stork bitten by a crocodile; 7 events, 26 s total). As described in Jacoby et al. (2016), these conditions were identified and coded by 5 independent coders. Participants watched the film passively. (The localizer is available at http://saxelab.mit.edu/theory-mind-and-pain-matrix-localizer-movie-viewing-experiment; the Partly Cloudy short film itself must be purchased from Pixar Animation Studios.)

Jacoby et al. (2016) compared the activation patterns for the mental > pain contrast to those elicited by the verbal ToM contrast described above and found them to be similar in individual participants. Because ToM selectivity entails that the mental condition should elicit a stronger response than any of the other conditions (physical, social, and pain; e.g. Saxe and Powell 2006), here, we characterize our critical networks with respect to all 3 contrasts (mental > physical, mental > social, and mental > pain).

fMRI data acquisition, preprocessing, and first-level modeling

Data acquisition

Whole-brain structural and functional data were collected on a whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 176 axial slices with 1-mm isotropic voxels (repetition time [TR] = 2,530 ms; echo time [TE] = 3.48 ms). Functional, blood oxygenation level-dependent (BOLD) data were acquired using an EPI sequence with a 90° flip angle and using GRAPPA with an acceleration factor of 2; the following parameters were used: 31 4.4-mm thick near-axial slices acquired in an interleaved order (with 10% distance factor), with an in-plane resolution of 2.1 mm × 2.1 mm; FoV in the phase encoding (A ≫ P) direction 200 mm and matrix size 96 × 96 voxels; TR = 2,000 ms and TE = 30 ms. The first 10 s of each run were excluded to allow for steady-state magnetization.

Preprocessing

fMRI data were analyzed using SPM12 (release 7487), CONN EvLab module (release 19b) and other custom MATLAB scripts. Each participant's functional and structural data were converted from DICOM to NIFTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session (Friston et al. 1995). Potential outlier scans were identified from the resulting subject-motion estimates as well as from BOLD signal indicators using default thresholds in the CONN preprocessing pipeline (5 SD above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm, Nieto-Castanon 2020). Functional and structural data were independently normalized into a common space (the Montreal Neurological Institute [MNI] template; IXI549Space) using the SPM12 unified segmentation and normalization procedure (Ashburner and Friston 2005) with a reference functional image computed as the mean functional data after realignment across all timepoints omitting outlier scans. The output data were resampled to a common bounding box between MNI-space coordinates (−90, −126, −72) and (90, 90, 108) using 2-mm isotropic voxels and fourth-order spline interpolation for the functional data and 1-mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were then smoothed spatially using spatial convolution with a 4-mm FWHM Gaussian kernel.

First-level modeling

For both the language localizer task and the ToM localizer tasks, effects were estimated using a General Linear Model (GLM) in which each experimental condition was modeled with a boxcar function convolved with the canonical hemodynamic response function (HRF) (fixation was modeled implicitly such that all timepoints that did not correspond to 1 of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal timeseries were accounted for by a combination of high-pass filtering with a 128 s cutoff

and whitening using an AR(0.2) model (first-order autoregressive model linearized around the coefficient $a = 0.2$) to approximate the observed covariance of the functional data in the context of Restricted Maximum Likelihood estimation. In addition to main condition effects, other model parameters in the GLM design included first-order temporal derivatives for each condition (included to model variability in the HRF delays) as well as nuisance regressors controlling for the effect of slow linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal.

Definition of the language and ToM functional regions of interest

For each localizer, we defined a set of functional regions of interest (fROIs) using group-constrained, participant-specific localization (Fedorenko et al. 2010). For the core language fROIs, each individual map for the "sentences > nonwords" contrast was intersected with a set of 5 binary masks. These masks (available at OSF: https://osf.io/bzwm8/) were derived from a probabilistic activation overlap map for the same contrast in a large independent set of participants ($n = 220$) using watershed parcellation, as described in Fedorenko et al. (2010) for a smaller set of participants. These masks included 3 in the left frontal cortex—in the inferior and middle frontal gyri—and 2 in the left temporal cortex (Fig. 3). Within each mask, a participant-specific language fROI was defined as the top 10% of voxels with the highest *t*-values for the localizer contrast (see Lipkin et al. 2022 for evidence that the fROIs defined in this way are similar to the fROIs defined based on a fixed statistical significance threshold). In addition, we defined peripheral language fROIs: 1 in the left angular gyrus (using a mask derived in the same way as the masks for the core language areas) and 6 in the RH. Following e.g. Paunov et al. (2019), we defined masks for RH homotopes of core language areas and the RH AngG fROI by mirror-projecting the LH masks onto the RH and selecting the top 10% of most localizer-responsive voxels within these. In this way, the hemispheric symmetry only applies at the level of the masks; the particular sets of voxels selected within the LH vs. RH masks were free to differ within these symmetrical masks. Note that we distinguish between (i) core LH language areas and "peripheral" areas consisting of (ii) the RH language homotopes and (iii) the bilateral angular gyri (see Rationale for our functional networks approach section for our functional networks approach).

For the ToM fROIs, each individual map for the false belief > false photo contrast from the verbal ToM localizer was intersected with a set of 10 binary masks (5 in each hemisphere). These masks (available at OSF: https://osf.io/bzwm8/) were derived from a random effects map for the same contrast in a large independent set of 462 participants (Dufour et al. 2013). These masks included bilateral regions in temporoparietal junction, left and right precuneus/posterior cingulate cortex, and left and right dorsal, middle, and ventral medial prefrontal cortex (Fig. 2). Within each mask, a participant-specific ToM fROI was defined as the top 10% of voxels with the highest *t*-values for the localizer contrast.

Note that although Jacoby et al. (2016) report evidence that areas in the bilateral STS respond to both verbal and nonverbal ToM contrasts, these areas are not typically considered as part of the ToM network (e.g. Isik et al. 2017), and we therefore did not include them in our set of ToM masks. This choice does not undermine our core claims, however, because our question is whether language areas are also involved in ToM reasoning (with prior report of spatial overlap serving as motivation for considering this question). Because our language network masks encompass the STS, they should capture the key areas within the STS that respond to language, including any such area that also responds to ToM. Thus, our analyses are conservative with respect to the proposition that language and ToM functions overlap in STS: Any such overlap should contribute to a ToM-like profile for the language-responsive STS areas when evaluated on a ToM task (which is not what we find).

Validation of the language and ToM fROIs

To ensure that the language and ToM fROIs behave as expected (i.e. language fROIs show a reliably greater response to the sentences condition compared to the nonwords condition; ToM fROIs show a reliably greater response to the false belief condition than the false photo condition), we used an across-runs cross-validation procedure (e.g. Nieto-Castañón and Fedorenko 2012). In this analysis, the first run of the localizer was used to define the fROIs, and the second run to estimate the responses to the localizer conditions (in percent BOLD signal change [PSC]), ensuring independence (e.g. Kriegeskorte et al. 2009); then the second run was used to define the fROIs, and the first run was used to estimate the responses; finally, the extracted magnitudes were averaged across the 2 runs to derive a single response magnitude for each of the localizer conditions. Statistical analyses were performed on these extracted PSC values. Note that this cross-validation approach was only used for validation and effect estimation of the localizer contrasts themselves, to ensure independence. For analyses that use a localizer to constrain the estimation of a different effect (e.g. analyses that used the verbal ToM localizer to define fROIs for estimating nonverbal ToM contrasts), the contrast maps from different localizer runs were combined.

Rationale for our functional networks approach

Our focus on functional networks draws on decades of research in systems neuroscience using coactivation of distributed brain areas as evidence of functional integration between those areas (Friston 2011; Hutchison et al. 2013; Glasser et al. 2016). Our analyses assume the existence of (i) a functionally integrated "core" LH language network; (ii) a functional dissociation between this core language network and what we call its "periphery" (e.g. Chai et al. 2016), which includes the RH language homotopes and the language areas in the bilateral angular gyri; and (iii) a functional dissociation within the periphery of the language network—with stronger functional integration within the RH language network and between the bilateral language areas in the angular gyri than between these sets of regions—which justifies our distinct treatment of the RH language homotopes and the language areas in the bilateral angular gyri.

These assumptions are based on robust and highly replicable patterns of inter-region correlations (IRCs) during naturalistic cognition paradigms. Figure 2A visualizes these patterns using data from a large-scale study across 86 native speakers of 45 typologically diverse languages (Malik-Moraleda et al. 2022; see Blank et al. 2014; Paunov et al. 2019 for replications). Figure 2B reports the average IRC within and between groups of language fROIs drawn from the lower triangle of the correlation matrix shown in Fig. 2A. Using this dataset, we statistically test Fisher-transformed average correlations within and between these fROI groups. Within group correlations are tested against 0 using bootstrap resampling across participants and region-region pairs. Between-group differences in the correlation are tested against 0 using a permutation test of the absolute difference in Fisher-transformed average correlation. All tests use 10,000 resampling

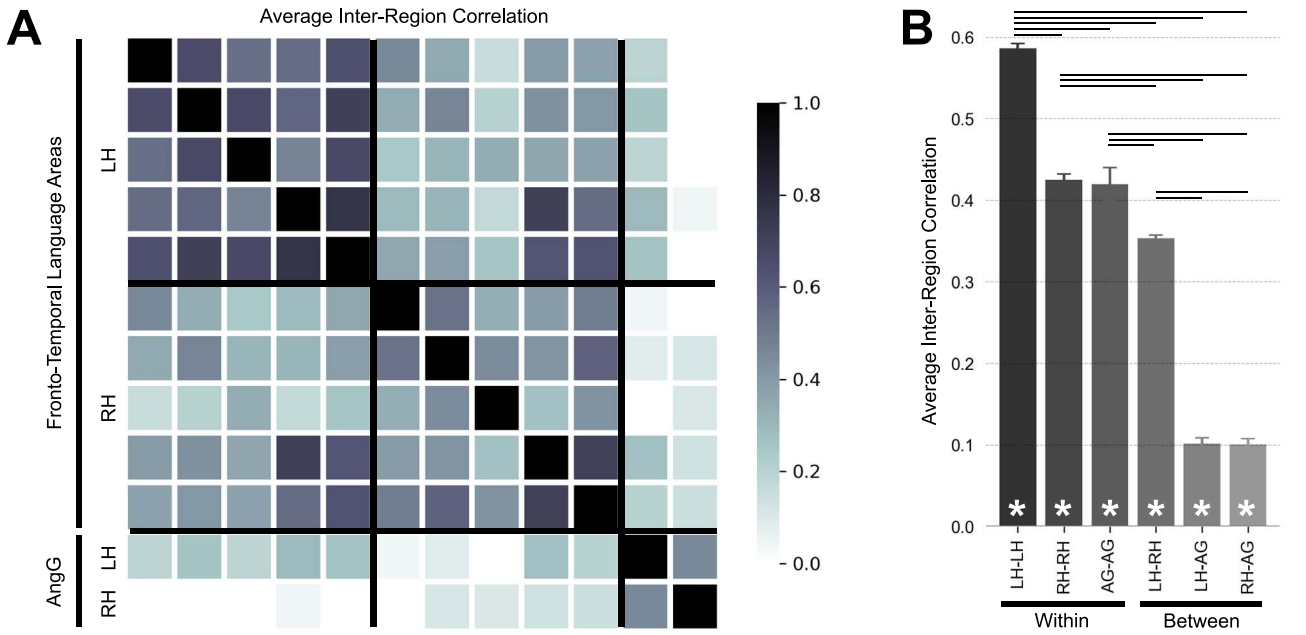

Fig. 2. A) Average IRCs across bilateral language-responsive areas in the inferior frontal cortex, temporal cortex, and angular gyrus during story comprehension and rest (averaging across these 2 naturalistic conditions), which were computed using publicly available data from Malik-Moraleda et al. (2022). The LH fronto-temporal language regions are strongly inter-correlated. The RH fronto-temporal language areas and language areas in the bilateral angular gyri are also internally correlated, although to a lesser extent than the LH language areas. B) Average (off-diagonal) IRCs within (first 3 bars) and between (last 3 bars) sets of language areas (i.e. LH language areas [LH], RH language homotopes [RH], and language areas in the angular gyri [AG]). Error bars show standard errors of the mean across participants and region-region pairs. Horizontal bars show significant differences between average IRCs by permutation test (all $P < 0.0001$). All within- and between-set average IRCs are significantly >0 by bootstrap test (all $P < 0.0001$), but critically, the correlations within each of these 3 sets are higher than they are between them.

iterations. As shown, we observe strong IRCs among the LH language regions (mean Fisher's $z = 0.75$, $P < 0.0001$), among the RH language regions (mean Fisher's $z = 0.51$, $P < 0.0001$), and between the language regions in the bilateral angular gyri (mean Fisher's $z = 0.49$, $P < 0.0001$). The IRCs among the LH language regions are stronger than among the RH language regions (mean difference in Fisher's $z = 0.24$, $P < 0.0001$) or between the language regions in the angular gyri (mean difference in Fisher's $z = 0.25$, $P < 0.0001$). The IRCs for the RH language regions and for the language regions in the angular gyri do not differ from each other (mean difference in Fisher's $z = 0.02$, $P = 0.57$).

Importantly, the IRCs "within" each of the 3 sets of language areas (LH, RH, and AG) are stronger than "between" them, evidencing dissociations. In particular, the LH language regions are more correlated with each other than they are with either the RH language regions (mean difference in Fisher's $z = 0.33$, $P < 0.0001$) or with the language regions in the angular gyri (mean difference in Fisher's $z = 0.63$, $P < 0.0001$). Likewise, the RH language regions are more internally correlated than they are with either the LH language regions (mean difference in Fisher's $z = 0.10$, $P < 0.0001$) or with the language regions in the angular gyri (mean difference in Fisher's $z = 0.40$, $P < 0.0001$). Finally, the language regions in the angular gyri are more correlated with each other than they are with either the LH language regions (mean difference in Fisher's $z = 0.38$, $P < 0.0001$) or with the RH language regions (mean difference in Fisher's $z = 0.38$, $P < 0.0001$).

Together, these results support (i) our chosen partitioning of different language areas into larger functional groups for analysis (stronger within-group than between-group IRCs); (ii) our particular focus on the 5 LH frontal and temporal language areas as the "core" network that supports human language processing (strongest IRCs); and (iii) our characterization of the RH language homotopes and the language areas in the angular gyri as "peripheral" (weaker IRCs). For validation analyses supporting our functional localization approach, see Supplementary Information.

That said, there are necessary choice points in our analysis which are motivated as much by simplicity and citation precedent (see recent work from the Fedorenko group, e.g. Blank, Balewski, et al. 2016a; Paunov et al. 2019, 2022; Shain et al. 2022) as by empirical results. First, the granularity of our particular parcellation of brain tissue into regions could, in principle, range from a single whole-brain mask, to lobar masks, to masks that correspond to different regions within each lobe (what we do here), and to masks for even smaller subregions (in the limit, single voxels). The functional localization paradigm does not dictate a specific choice of granularity but rather imposes a trade-off: The larger the masks, the greater the potential to capture interindividual variation in the spatial distribution of function, but the poorer the ability to resolve network-internal functional differences. In the limit of single-voxel "masks" at the highest extreme of the granularity spectrum, there is no ability to capture interindividual variation in the spatial distribution of function, and the distinction between participant-specific functional localization and voxel-wise group averaging disappears (cf. Nieto-Castañón and Fedorenko 2012 for an approach that allows incorporating functional localization into voxel-wise analyses; "Subject-specific localizers in the context of whole-brain voxel-based analyses" section). Our sets of masks reflect the spatial distribution of responsiveness to language and ToM contrasts in large cohorts of independent participants (see Fedorenko et al. 2010 and Julian et al. 2012 for details of deriving the masks) and have been successfully used to investigate the numerous aspects of brain function (see General approach section). The use of the same masks across studies ensures continuity and easier

across-study comparisons—a foundation of the cumulative scientific enterprise.

Second, our sets of masks could, in principle, be expanded to include additional language- and/or ToM-responsive areas, including e.g. parts of the hippocampus (for language, Blank, Balewski, et al. 2016a; Blank, Duff, et al. 2016b), STS (for ToM, Deen et al. 2015), or cerebellum (for both, LeBel et al. 2021). We have chosen to focus on the subset of language-responsive regions that cover the lateral frontal and temporal cortical surfaces, which has been the focus of most past studies of language processing. Consequently, our conclusions apply to this subset of language-responsive areas, and future work may investigate the role in ToM of other components of the extended language network.

Statistical analysis

We use PSC values derived from the localizer tasks to define dependent variables in linear mixed effects models in lme4 (Bates et al. 2015) when examining entire networks, with random effects for Participant and fROI, or in simple linear models when examining the fROIs separately. When examining the fROIs separately, reported *P* values are adjusted for false discovery rate (Benjamini and Yekutieli 2001) over the number of fROIs in the network. When examining the periphery of the language network, we treat the RH homotopic areas ($n = 5$) and bilateral AngG areas ($n = 2$) separately and correct for the number of fROIs within each set.

Units of analysis

We conduct fROI-based analyses of critical effects in the core LH language network (comprised of the LIFGorb, LIFG, LMFG, LAntTemp, and LPostTemp fROIs) and the bilateral ToM network (comprised of bilateral temporo-parietal junction [TPJ], DMPFC, MMPFC, VMPFC, and PC fROIs). We additionally conduct parallel analyses at the level of each individual language and ToM fROI. Finally, as described above, we analyze 2 key components of the "periphery" of the language network (Chai et al. 2016). This periphery is comprised of the RH homotopes of the LH language fROIs and fROIs in the bilateral angular gyrus (AngG). These areas have previously been implicated in social cognition. In particular, the RH homotopes of the language areas have been argued to support social processing (e.g. Rajimehr et al. 2022), and the bilateral angular gyrus (AngG) is 1 of the key areas where overlap between language and ToM contrasts has been previously reported (Deen et al. 2015).

Because only a subset of our participants (48/151) completed the nonverbal ToM task, analyses that involve nonverbal contrasts are restricted to those participants that completed all 3 tasks (language localizer, verbal ToM localizer, and nonverbal ToM localizer). Otherwise, analyses included the 149 participants with 2 ToM localizer runs.

Main analyses

Our localizers provide the following key conditions:

- **Language localizer:** Sentences and Nonwords
- **Verbal ToM localizer:** False Belief and False Photo
- **Nonverbal ToM localizer:** Mental, Physical, Social, and Pain

In language areas, sentences should elicit a larger response than nonwords. In ToM areas, false belief stories should elicit a larger response than false photo stories, and video segments with mental content should elicit a larger response than video segments with physical content (as well as segments that depict physical pain [Jacoby et al. 2016], and, to a lesser extent, segments that depict social interactions). Critically, if the language areas support some aspects of ToM, they should also show a false belief > false photo contrast and a stronger response to the mental condition than the nonmental (physical, social, and pain) conditions.

To test a contrast, we use the following linear mixed effects model, where "PSC" reflects the PSC (relative to the fixation baseline) associated with a given condition in a specific participant and fROI, "Contrast" is a binary indicator variable indexing which condition of the critical contrast is reflected by the PSC (e.g. whether the PSC measures the response to nonwords or sentences in the sentences > nonwords contrast), and the critical bolded variable is evaluated with a likelihood ratio test:

$$\text{PSC} \sim 1 + \textbf{Contrast} + (1 \mid \text{Participant}) + (1 \mid \text{fROI})$$

A fixed-effects only variant of this model is used for tests in individual fROIs (Participant cannot be a random effect in this design because each datapoint has a unique participant):

$$\text{PSC} \sim 1 + \textbf{Contrast}$$

Note that we could have included by-participant and by-fROI random slopes for "Contrast" in the network model and a by-participant random intercept in the individual fROI model, but we found that doing so led to frequent problems with model identification in critical tests (nonconvergence or singular fits).

Linguistic analyses

To better understand any linguistic determinants of verbal ToM effects in language regions, we analyzed the verbal ToM materials in terms of linguistic properties that are known, based on past behavioral and neural findings, to modulate language network activity. If the false belief conditions differ systematically from the false photo conditions in ≥1 of these dimensions, these differences could account for the differences in language network activation. We considered the following linguistic predictors:

- **Num Words:** The number of words in an item. Language network activity has previously been associated with the length of linguistically coherent spans (e.g. Pallier et al. 2011; Fedorenko et al. 2016).
- **Num Sents:** The number of sentences in an item, which may modulate language network activity via sentence wrap-up processes (e.g. Just and Carpenter 1980; Rayner et al. 2000).
- **Constituent End:** Whether a word terminates a syntactic constituent in a hand-corrected phrase structure tree. Constituent boundaries may modulate language network activity via constituent wrap-up processes (Nelson et al. 2017).
- **Integration Cost:** A measure of working memory retrieval difficulty. Integration cost is posited by the Dependency Locality Theory (DLT; Gibson 2000) as an account of word-by-word variation in the difficulty of building linguistic representations in working memory. Here, we use a variant of DLT integration cost which has been associated with language network activity in prior work (Shain et al. 2022).
- **Unigram Surprisal:** A measure of word frequency, specifically: The negative log of a word's marginal probability according to a unigram KenLM language models (Heafield et al. 2013) trained on the Gigaword 3 corpus (Graff et al. 2007). Stronger language network activation has been associated with less frequent words (higher unigram surprisal, e.g. Schuster et al. 2016).
- **5-gram Surprisal:** A measure of word predictability, specifically: The negative log probability of a word in context according to a 5-gram KenLM language model (Heafield et al. 2013)

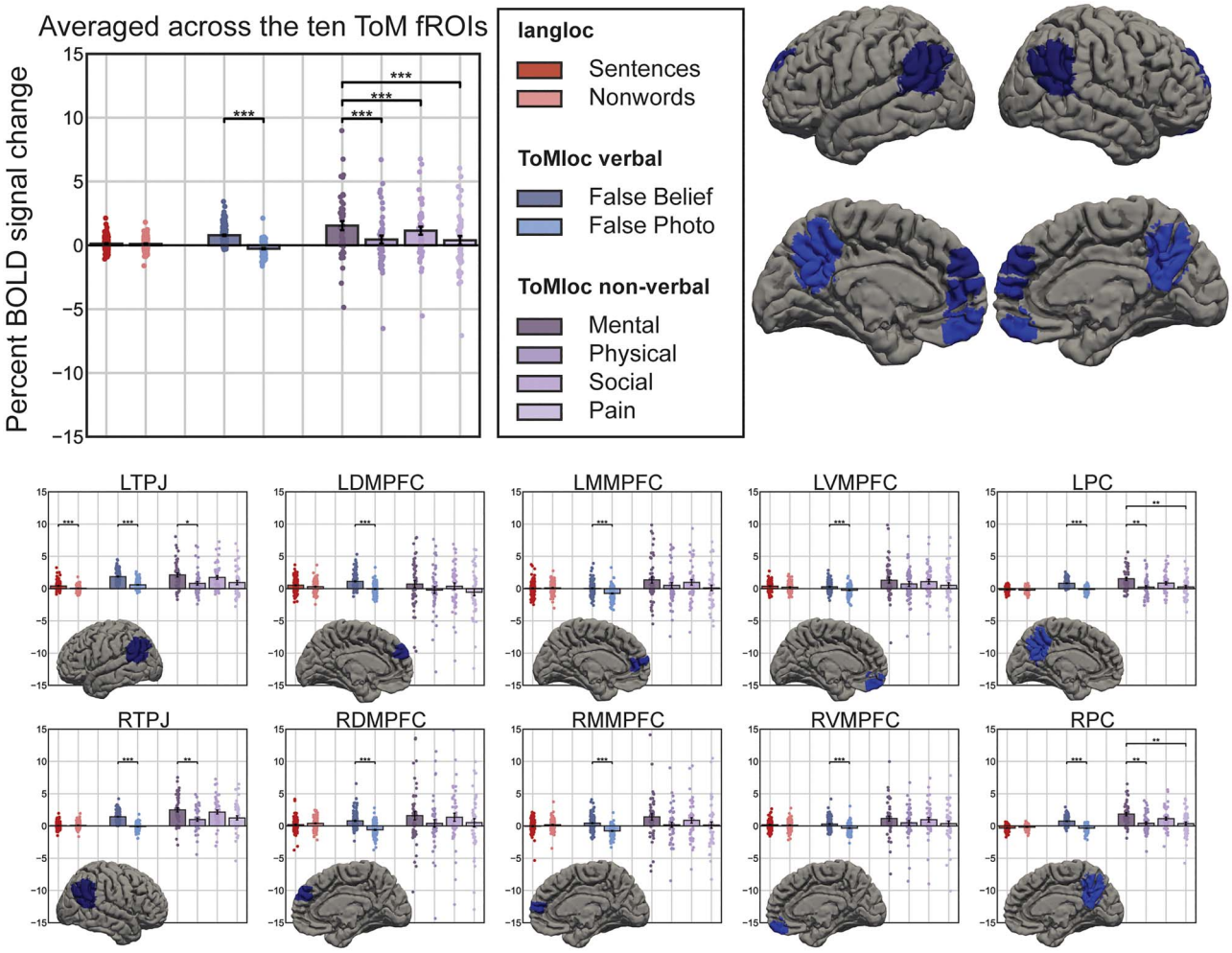

Fig. 3. Responses to the conditions of the language localizer and the verbal and nonverbal ToM localizers in the ToM network. The ToM network is not sensitive to the language contrast. ToM network activity increases in the presence of mental content whether mediated verbally (false belief > false photo) or nonverbally (mental > physical, mental > social, mental > pain). This overall pattern of results also holds qualitatively within each of the 10 regions of the ToM network.

trained on the Gigaword 3 corpus (Graff et al. 2007). Stronger language network activation has been associated with less predictable words (higher 5-gram surprisal, e.g. Lopopolo et al. 2017; Shain et al. 2020).

- **PCFG Surprisal:** A measure of word predictability, specifically: The negative log probability of a word in context according to a probabilistic context-free grammar (PCFG) parser (van Schijndel et al. 2013) trained on a generalized categorical grammar reannotation (Nguyen et al. 2012) of the Wall Street Journal portion of the Penn Treebank corpus (Marcus et al. 1993). PCFG and 5-gram Surprisal effects have been shown to be dissociable in the human language network (Shain et al. 2020).

Item-level values for Constituent End, Integration Cost, Unigram Surprisal, 5-gram Surprisal, and PCFG Surprisal were computed by averaging their respective values over all words in an item.

We ask whether controlling for these linguistic variables attenuates the false belief > false photo contrast in the language network. To investigate this question, we first regress each variable individually out of the item-wise PSCs in each language region of each participant. We then compute the change in the false belief > false photo contrast (i.e. the change in the difference between the average response to false belief items and the average response to false photo items) due to a linguistic control. For simplicity, we refer to the change due to linguistic feature *X* in the false belief > false photo effect as ΔToM.*X*. To test ΔToM.*X* for significance in a given functional network, we model it as the dependent variable in linear mixed effects models with the following structure, where the fixed intercept is evaluated with a likelihood ratio test:

$$\Delta\text{ToM.X} \sim \mathbf{1} + (1 \mid \text{Participant}) + (1 \mid \text{fROI})$$

Similarly, we also examine the combined effect of regressing out all control variables simultaneously. In individual fROIs, we test ΔToM.*X* with a 1-sample *t*-test.

Results

The ToM network shows both verbal and nonverbal ToM effects

Replicating prior work (e.g. Jacoby et al. 2016), our results show that functionally localized regions previously associated with ToM reasoning are significantly more activated in the

Table 1. Size and significance of key contrasts in the ToM network (overall) and each of its 10-component fROIs (fROI-level *P* values are FDR-corrected for 10 fROIs).

	Belief > photo		Mental > physical		Mental > social		Mental > pain	
	β	P	β	P	β	P	β	P
Overall	1.05	<0.001***	1.09	<0.001***	0.40	<0.001***	1.15	<0.001***
LTPJ	1.27	<0.001***	1.30	0.02*	0.39	1.00	1.18	0.07
LDMPFC	1.17	<0.001***	0.93	0.72	0.35	1.00	1.27	0.40
LMMPFC	0.79	<0.001***	0.83	0.72	0.36	1.00	1.22	0.38
LVMPFC	0.53	<0.001***	0.57	0.92	0.21	1.00	0.79	0.59
LPC	0.93	<0.001***	1.28	0.002**	0.67	0.69	1.25	0.005**
RTPJ	1.51	<0.001***	1.49	0.004**	0.33	1.00	1.24	0.07
RDMPFC	1.39	<0.001***	1.19	0.72	0.24	1.00	1.08	0.68
RMMPFC	1.21	<0.001***	1.23	0.34	0.53	1.00	1.25	0.38
RVMPFC	0.61	<0.001***	0.67	0.72	0.18	1.00	0.79	0.59
RPC	1.09	<0.001***	1.39	0.002**	0.69	0.69	1.46	0.003**

Numerical estimates (β) and network-wide significance tests show a selective response to mentalizing across verbal and nonverbal representation formats. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

presence of mental state content (Fig. 3) whether this content is delivered verbally (false belief > false photo) or nonverbally (mental > physical, mental > social, and mental > pain). The false belief > false photo contrast is significant in the network overall ($\beta = 1.05$, $P < 0.001$***) as well as in each individual ToM fROI (Table 1). The mental > physical, mental > social, and mental > pain contrasts are significant overall ($\beta = 1.09$, 0.40, and 1.15, respectively; all $P < 0.001$***) and numerically positive in each individual ToM fROI, achieving significance at the fROI level in bilateral PC (mental > physical and mental > pain) and bilateral TPJ (mental > physical) (Table 1). These results support a selective role for this network in the mentalizing aspects of ToM ("cognitive ToM"; Saxe and Powell 2006; Bruneau, Pluta, et al. 2012b). Note that we did not evaluate the language localizer contrast in the ToM network statistically because our localizer materials are not controlled for mental state content, but as can be seen in Fig. 2, responses to these conditions are generally low, with little difference between the sentences and nonwords conditions (see also Koster-Hale and Saxe 2013 and Deen et al. 2015, which show the lack of engagement of the ToM network for sentences devoid of mental/social content in the presence of robust responses to those stimuli in the language network).

The language network shows verbal but not nonverbal ToM effects

Replicating much prior work, the core LH language network (Fig. 4) shows a larger response to sentences over nonword lists (Fedorenko et al. 2010), and replicating Deen et al. (2015), to false belief items over false photo items in the verbal ToM localizer. Both of these contrasts are significant in the language network as a whole (sentences > nonwords: $\beta = 1.49$, $P < 0.001$***; false belief > false photo: $\beta = 0.58$, $P < 0.001$***) and in each individual language fROI (Table 2). The overall response to both conditions of the verbal ToM localizer is more similar to the response to sentences than to the response to nonwords (as expected, given that both ToM conditions are presented in coherent language), and indeed both verbal ToM conditions elicit a significantly larger response than the nonwords condition of the language localizer both in the language network as a whole (false belief > nonwords: $\beta = 1.78$, $P < 0.001$***; false photo > nonwords: $\beta = 1.20$, $P < 0.001$***) and in each individual language fROI.

However, the ToM effect is greatly attenuated when using a nonverbal (mental > physical, mental > social, mental > pain) contrast. Neither the language network as a whole nor any fROI within it registers a significant mental > physical effect ($\beta = 0.12$, $P = 0.29$). Furthermore, the language network is significantly less responsive to segments with mental content than to segments that depict nonmental social interactions (mental > social: $\beta = -0.25$, $P = 0.03$*; Table 2). The only ToM-like feature of the language network's response to the nonverbal ToM localizer is a greater response to segments that depict mental content than to segments that involve physical pain (mental > pain: $\beta = 0.29$, $P = 0.009$**). However, this contrast alone is insufficient to demonstrate ToM selectivity in the absence of selectivity for mentalizing over physical and social events. Thus, unlike the ToM network, we do not find evidence that the language network is selectively engaged by reasoning about the content of others' minds.

In addition, the overall response to the conditions of the nonverbal ToM localizer is more similar in magnitude to the response to nonwords than to the response to sentences, and indeed both nonverbal ToM conditions elicit a significantly smaller response than the sentences condition of the language localizer both in the language network as a whole (sentences > mental: $\beta = 1.33$, $P < 0.001$***; sentences > physical: $\beta = 1.44$, $P < 0.001$***) and in most individual language fROIs (Table 3).

Linguistic features explain verbal ToM effects in the language network

The fact that ToM effects only emerge in the language network for a verbal contrast (cf. the ToM network, where both verbal and nonverbal ToM contrasts elicit an effect) suggests that this effect may reflect linguistic differences between the false belief and false photo conditions of the verbal ToM localizer task rather than ToM reasoning. If true, this hypothesis predicts (i) that the false belief and false photo conditions will systematically differ in linguistic features that modulate language network activity independently of ToM and thus (ii) that controlling for the relevant features will attenuate the false belief > false photo effect in the language network. To test this hypothesis, we analyzed the effect of controlling for 7 independently motivated linguistic features on the size of the false belief > false photo contrast: Num Words, Num Sents, and item-wise averages of Constituent End, Integration Cost, Unigram Surprisal, 5-gram Surprisal, and Probabilistic Context-Free Grammar (PCFG) Surprisal, for definitions, see Linguistic analyses section). The

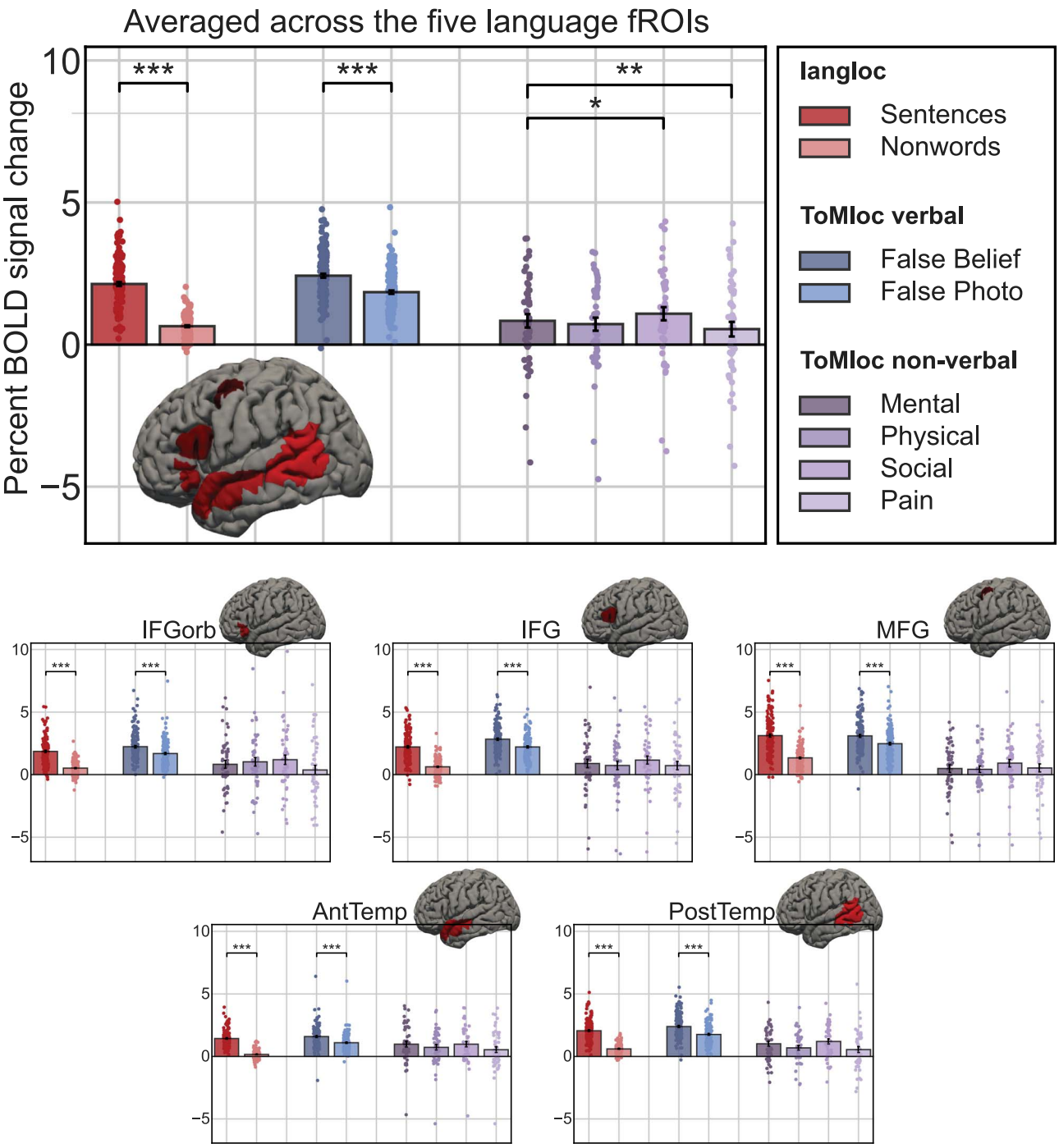

Fig. 4. Responses to the conditions of the language localizer and verbal and nonverbal ToM localizers in the language network. Replicating prior work, the language network shows a robust sentences > nonwords contrast. We also observe a false belief > false photo contrast in the verbal ToM task. However, the language network shows no significant mental > physical contrast and a negative mental > social contrast in the nonverbal ToM task, which is not consistent with ToM selectivity and suggests that the false belief > false photo contrast may be driven by linguistic differences between the 2 conditions. This overall pattern of results also holds within each of the 5 regions of the language network. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

distributions of these features in the false belief and false photo materials are visualized in Fig. 5A. As shown, false belief items are systematically higher in dimensions that are known to modulate language network activity, including Num Words*, Num Sents*, Integration Cost*, 5-gram Surprisal*, and PCFG Surprisal (stars indicate significant differences in a 2-sample t-test). These feature distributions are consistent with our hypothesis that the verbal ToM contrast is confounded with linguistic complexity.

To test the hypothesis directly, we analyzed the impact of controlling for each linguistic feature individually, as well as the impact of controlling for all seven linguistic features jointly, on the magnitude of the verbal ToM contrast in the language network (for statistical procedures, see Statistical analysis section). Results are plotted in Fig. 6. The predictors Num Words, Num Sents, Constituent End, Integration Cost, and 5-gram Surprisal significantly attenuate the verbal ToM contrast in the language network as a whole as well as in each fROI within it (Table 4). PCFG

Table 2. Size and significance of key contrasts in the language network (overall) and each of its 5-component fROIs (fROI-level *P* values are FDR-corrected).

	Sent > nonwd		Belief > photo		Mental > physical		Mental > social		Mental > pain	
	β	*P*	*β*	*P*	*β*	*P*	*β*	*P*	*β*	*P*
Overall	1.49	<0.001***	0.58	<0.001***	0.12	0.29	−0.25	0.03*	0.29	0.009**
LIFGop	1.34	<0.001***	0.54	<0.001***	−0.20	1.00	−0.36	1.00	0.45	1.00
LIFG	1.58	<0.001***	0.62	<0.001***	0.15	1.00	−0.28	1.00	0.16	1.00
LMFG	1.78	<0.001***	0.62	<0.001***	0.05	1.00	−0.43	1.00	−0.05	1.00
LAntTemp	1.27	<0.001***	0.49	<0.001***	0.25	1.00	0.00	1.00	0.44	0.98
LPostTemp	1.45	<0.001***	0.63	<0.001***	0.33	1.00	−0.18	1.00	0.47	0.98

Numerical estimates (β) and network-wide significance tests show a selective response to language (sentence > nonword) and verbal theory of mind (ToM; Belief > Photo) but no robust response to nonverbal ToM (mental > physical) or selectivity for mentalizing over social interactions (mental < social). *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

Table 3. Size and significance of contrasts between sentences and mental/physical conditions of the nonverbal ToM localizer in the language network (overall) and each of its 5-component fROIs (fROI-level *P* values are FDR-corrected).

	Sent > mental		Sent > physical	
	β	*P*	*β*	*P*
Overall	1.33	<0.001***	1.44	<0.001***
LIFGop	1.07	0.01*	0.87	0.05
LIFG	1.24	0.005**	1.39	0.002**
LMFG	2.75	<0.001***	2.80	<0.001***
LAntTemp	0.49	0.11	0.74	0.006**
LpostTemp	1.09	<0.001***	1.42	<0.001***

Responses to sentences are significantly larger in the language network as a whole (overall) and in most of its 5-component fROIs. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

Table 4. Effects of controlling for linguistic variables in the core language network (fROI-level *P* values are FDR-corrected).

	Num words		Num Sents		Const end		Int cost		Unigram		5-gram		PCFG		All	
	β	*P*	*β*	*P*	*β*	*P*	*β*	*P*	*β*	*P*	*β*	*P*	*β*	*P*	*β*	*P*
Overall	−0.20	<0.001***	−0.13	<0.001***	−0.003	0.002**	−0.12	<0.001***	−0.003	0.33	−0.11	<0.001***	−0.02	0.02*	−0.49	<0.001***
LIFGop	−0.18	<0.001***	−0.13	<0.001***	−0.005	<0.001***	−0.13	<0.001***	−0.004	1.00	−0.11	<0.001***	−0.02	0.28	−0.46	<0.001***
LIFG	−0.23	<0.001***	−0.16	<0.001***	−0.005	<0.001***	−0.13	<0.001***	−0.005	1.00	−0.11	<0.001***	−0.02	0.12	−0.52	<0.001***
LMFG	−0.22	<0.001***	−0.13	<0.001***	−0.002	0.02*	−0.13	<0.001***	−0.006	1.00	−0.13	<0.001***	−0.04	0.001**	−0.55	<0.001***
LantTemp	−0.17	<0.001***	−0.10	<0.001***	−0.003	<0.001***	−0.08	<0.001***	0.001	1.00	−0.08	<0.001***	−0.01	0.45	−0.40	<0.001***
LpostTemp	−0.21	<0.001***	−0.14	<0.001***	−0.003	<0.001***	−0.11	<0.001***	−0.001	1.00	−0.12	<0.001***	−0.02	0.02*	−0.54	<0.001***

Effect estimates (β) represent the change in the language network's response to the verbal theory of mind (ToM) contrast (false belief > false photo) due to controlling for a linguistic variable. Most variables we considered significantly attenuate the ToM contrast in the language network as a whole and in most of its 5-component fROIs. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

Surprisal significantly attenuates the ToM contrast in the entire network as well as in LMFG and LPostTemp. Unigram Surprisal does not have a significant effect on the ToM contrast (see Shain 2019 for related findings). In addition, jointly controlling for all linguistic features attenuates the verbal ToM contrast by 0.49 ($P < 0.001$***) network-wide. Given that the network-wide false belief > false photo contrast is 0.58, this means that linguistic differences account for at least 84% of the verbal ToM effect in the language network, rendering its residualized effect size (0.09) comparable to that of the nonverbal mental > physical contrast in the language network (0.11). Although the attenuated verbal ToM effect remains significant in the LH language network overall and in each component fROI ($P < 0.001$***), our analysis only considered a handful of linguistic variables and therefore only provides a lower bound on the proportion of the verbal ToM contrast which is attributable to linguistic factors.

Deen et al. (2015) also reported language-ToM overlap using a broader but more carefully linguistically controlled contrast between an additional set of stories describing false beliefs and a set of stories describing physical changes with no mental state attribution. This overlap with the language system suggests that at least some language-responsive areas may be recruited for some social cognitive functions, including possibly ToM. Deen et al. (2015) fixed the number of sentences in each item at 3 and controlled for a diverse set of linguistic features: "number of words, mean syllables per word, Flesch reading ease, number of noun phrases, number of modifiers, number of higher-level constituents, number of words before the first verb, number of negations, and mean semantic frequency (log Celex frequency)" (Deen et al. 2015, p. 4598). Nonetheless, as shown in Fig. 5B, the false belief items differ statistically from the physical change items along every relevant dimension in our linguistic evaluation (number of sentences has no variance by construction, as noted above), and the false belief items are systematically higher in dimensions known to increase language network activity: They have higher average integration cost, involve less frequent words (higher unigram surprisal), and are less predictable on the basis of both word cooccurrences (5-gram surprisal) and syntactic structure (PCFG surprisal). Thus, differences between these conditions in language-selective areas are also plausibly driven

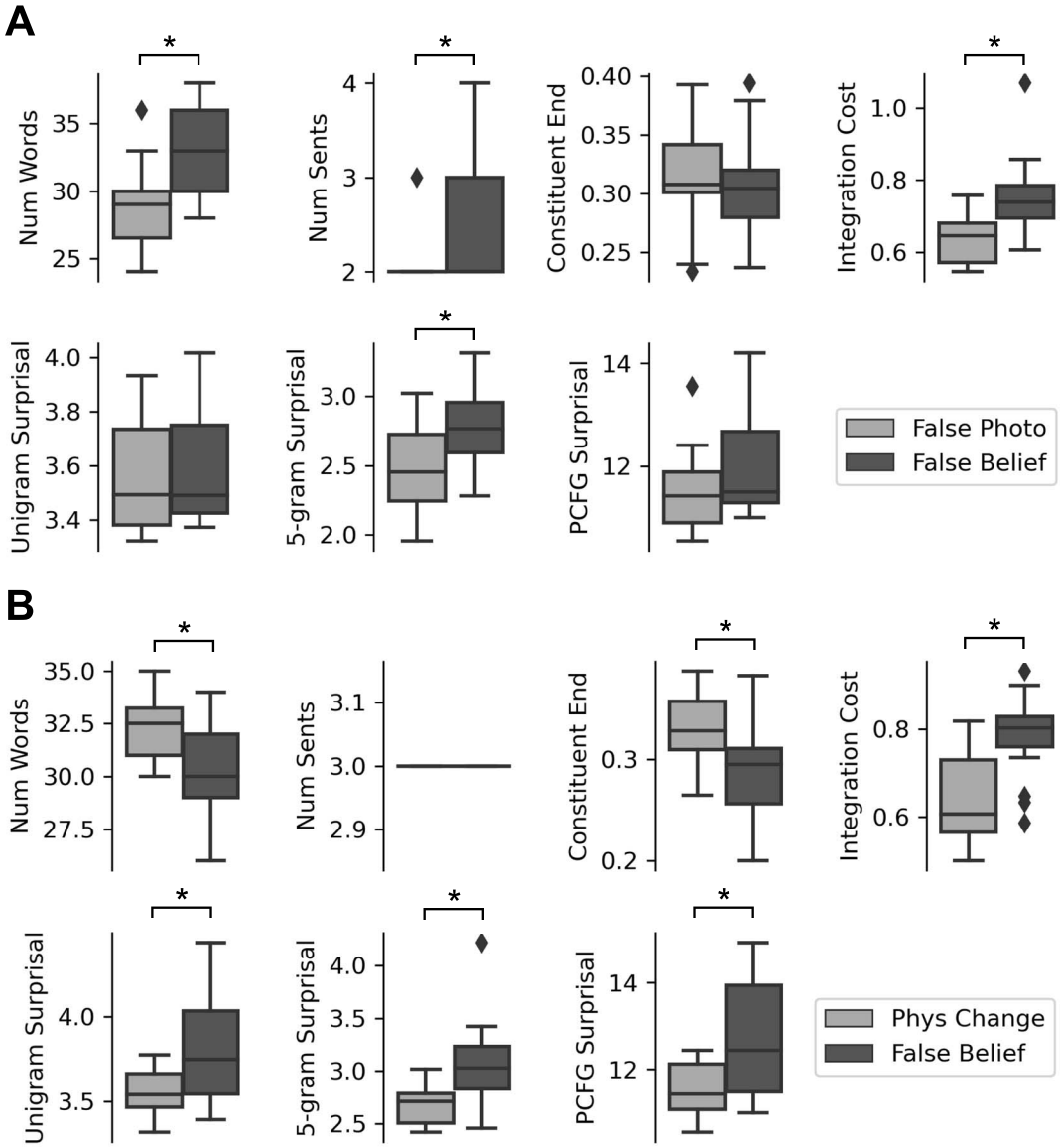

Fig. 5. A) Distribution of linguistic features in the false belief vs. false photo items of the verbal ToM localizer task. The false belief items are systematically higher in dimensions that are known to modulate language network activity, including number of words, number of sentences, integration cost, and 5-gram surprisal. B) Identical analyses for the false belief vs. physical change items (another set of verbal ToM-targeting materials) used in Deen et al. (2015). These items also differ significantly along dimensions known to modulate language network activity, especially integration cost, unigram surprisal, 5-gram surprisal, and PCFG surprisal. *: statistically significant.

by linguistic confounds despite considerable effort invested in matching along many linguistic features. The methodological upshot of this outcome is that linguistic matching of complex verbal stimuli is challenging, if not impossible, due to the myriad structural and statistical relationships that hold between words in language. For designs that seek to study the modulation of language-selective brain areas by content-related (semantic) contrasts, it may be necessary to avoid verbal stimuli, or at least to supplement verbal contrasts with nonverbal ones.

The language network's "periphery" may support broader social cognition

Even though the core LH language areas do not show evidence of supporting mental state attribution in our study, it has been argued that some regions in the periphery of the language processing system (Chai et al. 2016) are associated with ToM reasoning and/or social processing more generally. Here, we consider 2 candidate components of the language periphery: the RH homotopes of the LH core language regions and the language-responsive areas in the bilateral angular gyri. The function(s) of both of these components remains debated in the field (see Discussion).

Responses to the key conditions of all 3 localizers in the RH homotopes of the core language areas are plotted in Fig. 7. RH language regions show considerably less selectivity for language processing than their LH counterparts: Unlike in the LH, linguistic stimuli (the sentence condition of the language localizer, and the conditions of the verbal ToM localizer) elicit lower responses than the rich visual stimuli from the nonverbal ToM localizer (see Small et al. (2021) and Ivanova (in prep.-b) for additional evidence of lower selectivity of the RH language regions). In addition, unlike the core language network but similar to the ToM network, these

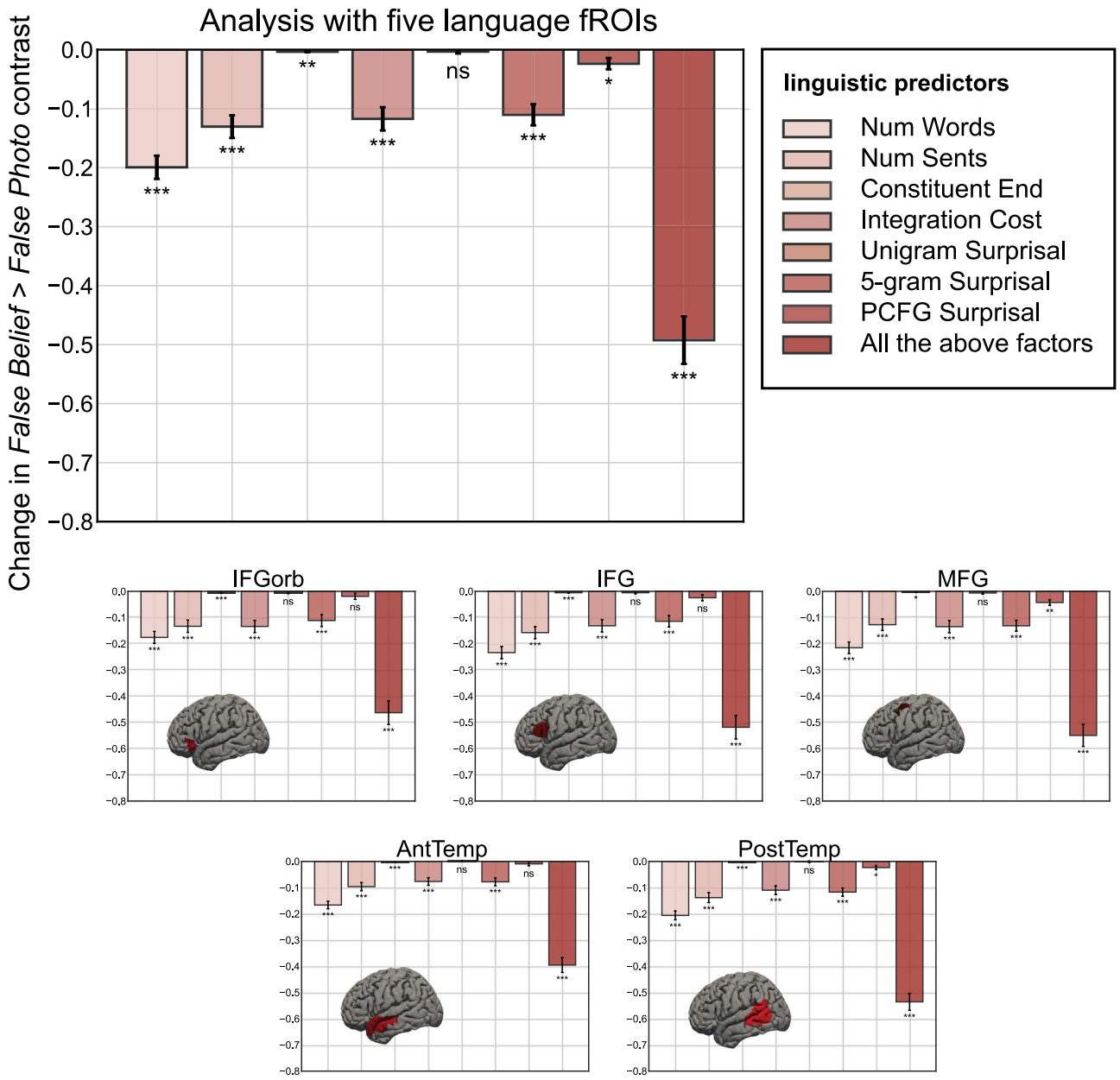

Fig. 6. Effects of controlling for linguistic features on the false belief > false photo contrast in the language network. Effects are universally negative, meaning that controlling for the variable systematically attenuates verbal ToM contrasts throughout the language network. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

RH language areas respond significantly to the false belief > false photo contrast ($\beta = 0.59$, $P < 0.001$***) and the mental > physical contrast ($\beta = 0.38$, $P = 0.001$**). However, unlike the ToM network, RH language areas are not selective for mentalizing segments relative to segments that depict physical pain (mental > pain: $\beta = -0.01$, $P = 0.94$), and they are significantly less responsive to mentalizing segments than to segments that depict nonmental forms of social interaction (mental > social: $\beta = -0.28$, $P = 0.019$*; Fig. 7, Table 5). Thus, any contribution of the RH language areas to social cognition is not restricted to cognitive ToM/mentalizing.

Responses to the conditions of all 3 localizers in the language-responsive areas in the angular gyrus (bilaterally) are plotted in Fig. 8. Like the ToM network, these areas respond significantly to the false belief > false photo contrasts ($\beta = 0.61$, $P < 0.001$***) and both the mental > physical ($\beta = 0.73$, $P < 0.001$***) and mental > pain contrasts ($\beta = 0.68$, $P < 0.001$***) of the nonverbal ToM localizer. However, unlike the ToM network, these areas do not show a mental > social effect ($\beta = 0.01$, $P = 0.96$; Fig. 8, Table 5). Thus, similar to what we observed for the RH language areas, any contribution of the language areas in the AngG to social cognition appears to be different from that of the ToM network in that it is not selective for ToM reasoning.

Figure 9 shows responses in the 4 sets of fROIs examined here to the 4 conditions of the nonverbal ToM localizer (mental, i.e. segments depicting mental state content; physical, i.e. segments depicting physical events; social, i.e. segments depicting non-mentalizing social interactions; and pain, i.e. segments depicting physical pain). Only the ToM network shows the characteristic profile of greater response to the mental condition than either the physical or social condition; in the language network and its periphery, the response to the social condition is at least as large as the response to the mental condition. The language network

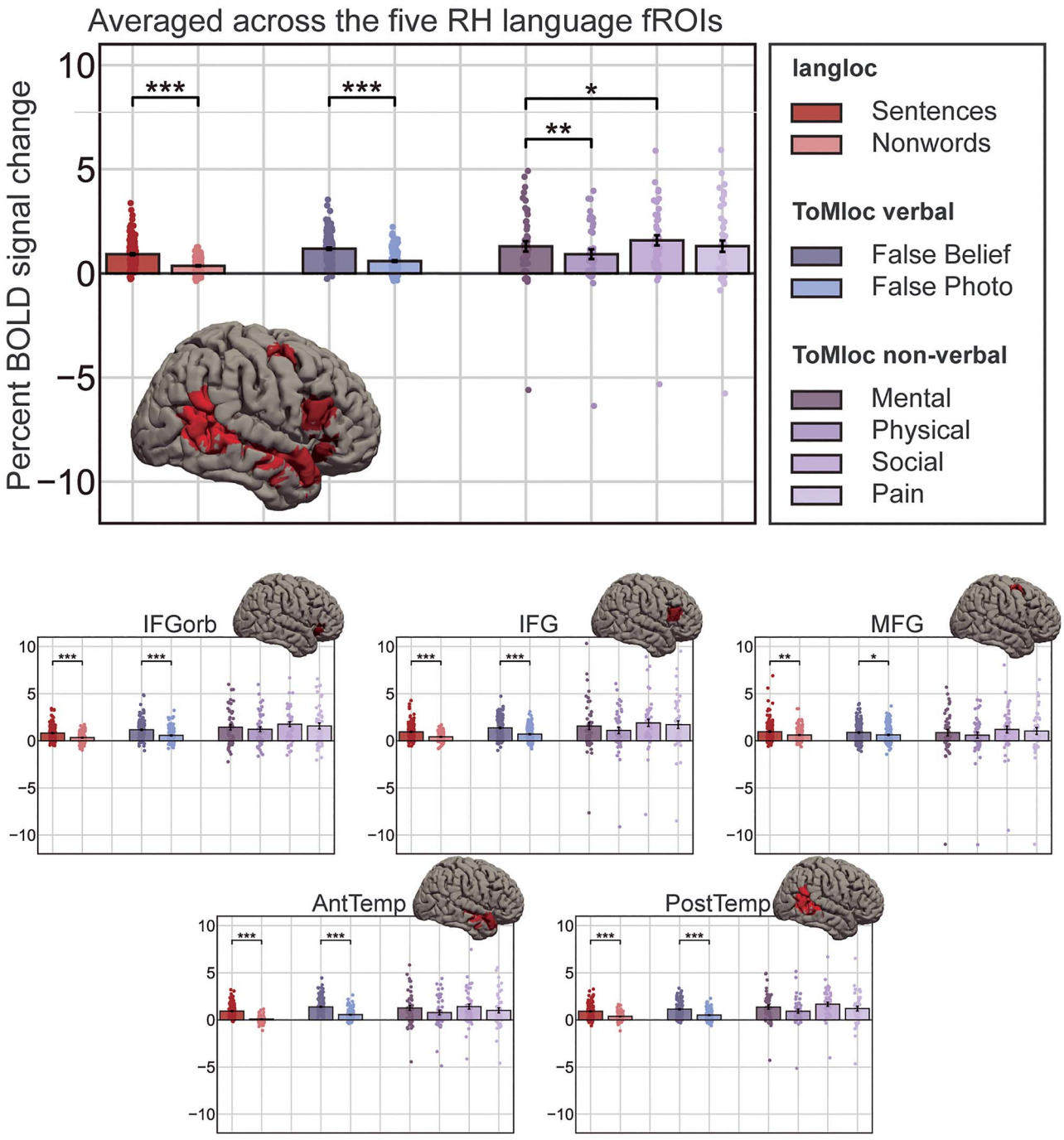

Fig. 7. Responses to the conditions of the language localizer and verbal and nonverbal ToM localizers in the RH homotopes of the language network. Replicating prior work, RH language regions show a robust sentences > nonwords contrast. However, unlike the core LH language network (Fig. 4), RH language regions show similarly strong responses to both the false belief > false photo contrast of the verbal ToM localizer and the mental > physical contrast of the nonverbal ToM localizer. This overall pattern of results also holds within each of the 5 RH homotopes of the core LH language network. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

response to all conditions in the task is lower than that of the other networks, likely due to the fact that this task is entirely nonverbal. The RH language homotopes and the angular gyri both show a stronger response to the mental and social conditions than to the "physical" and "pain" conditions, which is consistent with a broadly social function for these areas rather than a ToM-selective one.

Note that, based on much prior work (e.g. Saxe and Kanwisher 2003; Saxe and Powell 2006; Van Overwalle 2009; for review, see e.g. Saxe et al. 2004; Van Overwalle 2009), we are assuming the existence of the ToM network (i.e. a brain network that selectively supports ToM reasoning and that is spatially and functionally distinct from the language periphery), and we are merely reporting its responses for reference. Nonetheless, a surprising finding in, as shown Fig. 9, is that the ToM network itself shows a relatively large response to the "social" condition, unlike prior studies that reported clearer selectivity for the "mental" condition (Jacoby et al. 2016), suggesting that areas identified by the verbal ToM localizer may show a more generalized social response, albeit weaker than the response to mental state content. Since our present focus is

Table 5. Size and significance of key contrasts in the language periphery comprised of language-responsive fROIs in (i) the right hemisphere homotopes of core language areas and (ii) bilateral angular gyri (fROI-level *P* values are FDR-corrected).

	Sent > nonwd		Belief > photo		Mental > physical		Mental > social		Mental > pain	
	β	P	*β*	P	*β*	P	*β*	P	*β*	P
RH Overall	0.55	<0.001***	0.59	<0.001***	0.38	<0.001***	−0.29	0.02*	−0.01	0.94
RIFGop	0.49	<0.001***	0.60	<0.001***	0.22	1.00	−0.33	1.00	−0.14	1.00
RIFG	0.53	<0.001***	0.66	<0.001***	0.47	1.00	−0.34	1.00	−0.15	1.00
RMFG	0.36	0.001**	0.26	0.04*	0.26	1.00	−0.34	1.00	−0.18	1.00
RantTemp	0.85	<0.001***	0.82	<0.001***	0.48	1.00	−0.14	1.00	0.25	1.00
RpostTemp	0.55	<0.001***	0.62	<0.001***	0.44	1.00	−0.30	1.00	0.17	1.00
AngG Overall	0.46	<0.001***	0.61	<0.001***	0.73	<0.001***	0.01	0.96	0.68	0.001***
LangG	0.79	<0.001***	0.75	<0.001***	0.90	0.20	0.10	1.00	0.89	0.26
RangG	0.14	0.03*	0.47	<0.001***	0.55	0.21	−0.08	1.00	0.47	0.39

Like the language network, these areas respond to language (sentence > non-word), and like the theory of mind (ToM) network, they respond to verbal (belief > photo) and components of nonverbal (mental > physical) ToM. However, unlike the ToM network, they do not respond more to mentalizing relative to other forms of social interaction (null or negative mental > social effects), and the RH homotopic areas furthermore do not respond more to mentalizing relative to observing physical pain (mental > pain). These patterns are not consistent with a selective response to ToM but could be consistent with a more broadly social function in addition to language (but see Discussion for an alternative interpretation). *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

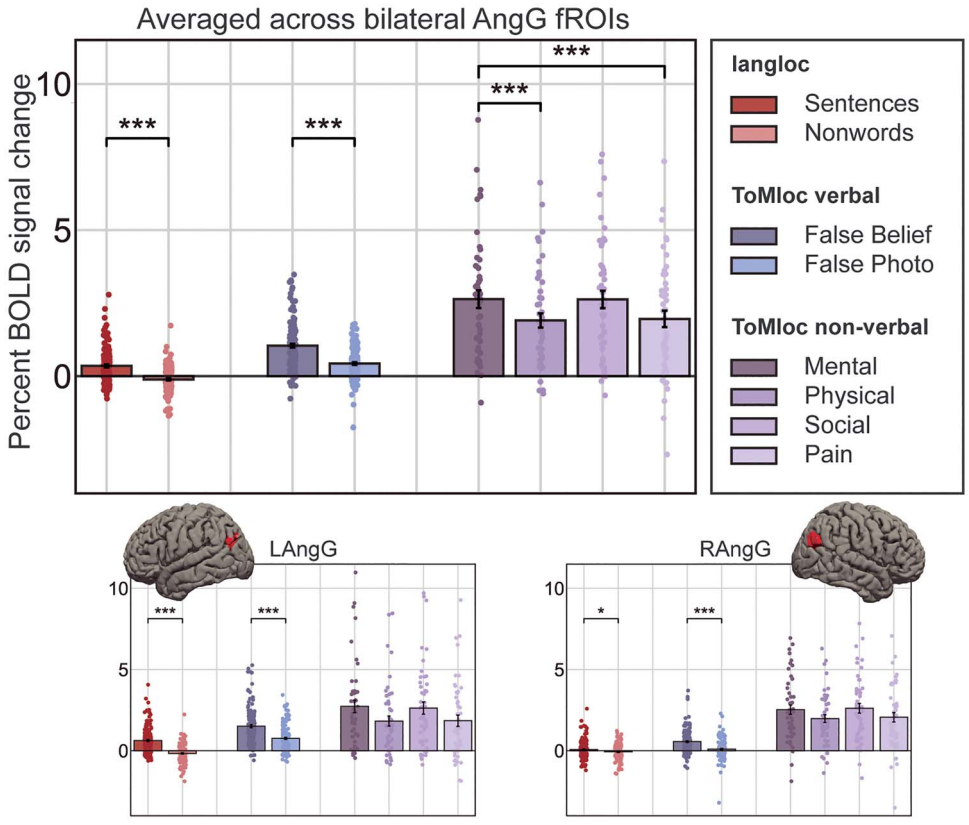

Fig. 8. Responses to the conditions of the language localizer and verbal and nonverbal ToM localizers in the bilateral angular gyri. Unlike the core LH language network (Fig. 4), these regions show similarly strong responses to both the false belief > false photo contrast of the verbal ToM localizer and the mental > physical contrast of the nonverbal ToM localizer. This overall pattern of results holds in both the LH and RH AngG fROIs. *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

on the functional role of language areas in ToM, rather than on previously established ToM areas, we leave further investigation of these questions (i.e. the contributions of the ToM network to social functions beyond mentalizing, and the relationship between the ToM network and the language periphery) to future work.

Discussion

Given the close functional relationship between language processing and thinking about others' thoughts (ToM), both developmentally (e.g. Astington and Jenkins 1999; Peterson and Siegal 2000; Hale and Tager-Flusberg 2003; Ruffman et al. 2003; Astington and Baird 2005; Slade and Ruffman 2005; Miller 2006; de Villiers and de Villiers 2014) and in adult language use (e.g. Grice 1975; Sperber and Wilson 1987; Winner et al. 1998; Champagne-Lavau and Joanette 2009; Roberts 2012), we asked whether the human language network, or some of its components, might additionally represent ToM information, as indicated by the recent findings from Deen et al. (2015). To investigate this question, we localized the language network in each participant in a

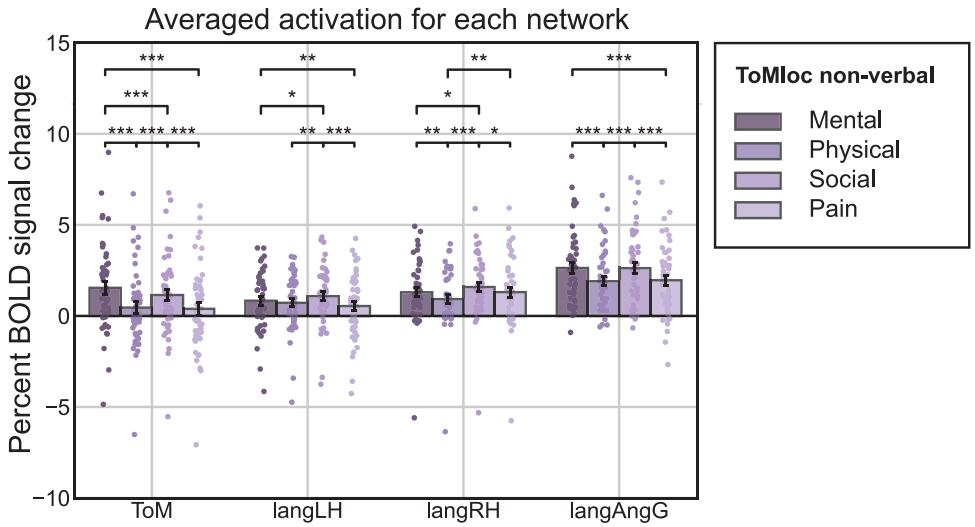

Fig. 9. Responses by network to the 4 conditions of the nonverbal ToM localizer (mental, physical, social, and pain). Whereas the ToM network is most strongly engaged by the mental condition, language regions in both hemispheres (langLH and langRH) are most strongly engaged by the social condition, and thus, in contrast to the ToM network, neither shows a selective response to mentalizing (mental > social). Language selective regions of the bilateral angular gyri (langAngG) are systematically more engaged by the social aspects of the task (mental/social > physical/pain) but likewise show no selectivity for mentalizing (mental > social). *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$.

large-scale fMRI study and evaluated the responses of these language areas to the established verbal ToM localizer task (based on the false belief > false photo contrast) and a more recently introduced nonverbal ToM task (based on mentalizing > nonmentalizing movie segments contrasts). Although the language network responds significantly to the verbal ToM contrast, it does not respond to the nonverbal ToM contrast, suggesting that the verbal ToM effect may be an artifact of linguistic differences between the conditions of the verbal ToM localizer. We confirmed this hypothesis by analyzing the verbal ToM materials with respect to linguistic features that are independently known to modulate activity in the language network. We showed that controlling for these features strongly attenuates the verbal ToM effect in language areas. It is thus likely that prior reports of language network activation in response to the verbal ToM contrast (Deen et al. 2015) were affected by these linguistic confounds.

In short, we do not find evidence that the core language areas are engaged in ToM reasoning. Nonetheless, both the nonverbal mental > physical and mental > pain contrasts and the linguistically residualized verbal false belief > false photo contrast are numerically positive in the language network as a whole as well as in some component fROIs. Furthermore, the mental > pain contrast is significant in the language network overall (albeit much smaller than the sentence > nonwords contrast), a result which is driven by the particularly low response of the language network to the pain condition and which warrants further investigation. We cannot rule out the possibility that these regions show a small increase in response to mentalizing that our current (relatively large) sample ($n = 149$ for verbal ToM and $n = 48$ for nonverbal ToM) lacks the power to detect. However, we have shown that any such effects are much smaller than effects of language processing and thus that the functional profile of these regions overwhelmingly favors language over ToM. Furthermore, the core language network does not show the characteristic selectivity for mentalizing; video segments depicting nonmentalizing social interactions induce a similar magnitude response to segments involving mentalizing. Thus, the language network shows neither a general response to ToM nor selectivity for ToM relative to other kinds of social processing, and prior evidence to the contrary (e.g. Deen et al. 2015) may have been driven by the linguistic confounds in the standard ToM localizer. Our results thus converge with recent findings from resting-state functional correlation analyses that independently identify a ToM-selective "default network B" (Braga and Buckner 2017; DiNicola et al. 2020) and show that this network is spatially distinct from the language network in individual brains (Braga et al. 2020).

In addition to our critical question about the involvement of core language areas in ToM processing, we additionally investigated the possible role in ToM of areas in the "periphery" of the language network (Chai et al. 2016) which have been implicated by prior work in ToM, social processing, or social/affective aspects of language processing: the RH homotopes of core language areas as well as language areas in the bilateral angular gyri.

The RH homotopes of the language regions respond to language contrasts, although generally less strongly (e.g. Fedorenko et al. 2010; Mahowald and Fedorenko 2016; Quillen et al. 2021; Lipkin et al. 2022; Martin et al. 2022). A number of claims have been made about the role of RH language regions in language processing and differential contributions of LH vs. RH language regions (e.g. Ross and Mesulam 1979; Bryan 1989; Bottini et al. 1994; Van Lancker 1997; Mitchell and Crow 2005; Lindell 2006; Beeman and Chiarello 2013). A common theme in this literature associates RH homotopes of language areas with the social, pragmatic, nonliteral, and/or affective aspects of speech processing and/or language comprehension (e.g. Van Lancker 1997; Mitchell and Crow 2005), including potentially a role in leveraging ToM for pragmatic inference (Kaplan et al. 1990). However, the empirical landscape is complex and ridden with controversy. Even the most common claim about the stronger role of the RH language areas, compared to the LH language areas, in nonliteral comprehension has been questioned (e.g. Lee and Dapretto 2006; Rapp et al. 2007, 2012; Paunov et al. 2019; Hauptman et al. 2022; see e.g. Calvo et al. 2019 for patient evidence). Based on the analyses of data from the Human Connectome Project (Van Essen et al. 2013), Rajimehr et al. (2022) recently argued that the primary function of these areas may be social rather than linguistic.

The function of the language-responsive areas in the left and right angular gyri also remains debated. A number of proposals have been put forward about the angular gyri in general (e.g. Farrer et al. 2008; Bonner et al. 2013; Price et al. 2015; Davis and Yee 2019; Humphreys et al. 2021) and their specific role in language processing (e.g. Thothathiri et al. 2012; Bemis and Pylkkänen 2013; Matchin et al. 2019; Branzi et al. 2021) as well as ToM processing (e.g. Saxe and Kanwisher 2003; Schurz et al. 2014, 2017). But, like other parts of the association cortex, the angular gyrus is a highly structurally and functionally heterogeneous area (e.g. Scholz et al. 2009; Uddin et al. 2010; Seghier 2013), which makes proposals about the entire angular gyrus difficult to evaluate. Of most relevance to the current investigation, Deen et al. (2015) observed some overlap between linguistic and ToM contrasts at the individual-participant level in the angular gyrus.

Unlike the core LH language areas, our analyses of the language periphery revealed a robust mental > physical contrast in the nonverbal ToM localizer, indicating that the language periphery indeed responds to mental content across representational formats (verbal and visual). However, unlike the ToM network, language fROIs in the RH and in the bilateral angular gyri respond as strongly, or even more strongly, to nonmentalizing social interactions, and the RH fROIs additionally register a strong response to observing others' physical pain. These response characteristics are not consistent with a selective response to ToM in the language network's periphery. They could be consistent with a broadly social function as proposed by e.g. Rajimehr et al. (2022) for the RH language homotopes. However, Rajimehr et al.'s claim is based on a single paradigm evaluated in a single (albeit large) dataset, and alternative explanations in terms of, for example, general visual semantic processing (e.g. Zaidel 1987; Joseph 1988) cannot be ruled out. Under such accounts, the somewhat stronger responses to social conditions would be explained by greater overall attention to social content rather than the processing of social content per se. Thus, more research is needed to understand the precise contribution of the RH language homotopes to semantic and specifically social cognition.

Conclusion

If the language network is not involved in making inferences about others' thoughts, how then do these inferences enter into language processing in order to inform rapid incremental sentence comprehension? (e.g. Shibata et al. 2010; Regel et al. 2011; Kaakinen et al. 2014). We hypothesize that this occurs via rapid feedback from the ToM network, which can then be used to inform interpretation. Although further research is needed to investigate this hypothesis, prior work has shown that the language and ToM networks show reliable functional correlations with each other over time during naturalistic cognition, which is consistent with information sharing (Paunov et al. 2019).

In conclusion, fMRI evidence supports a spatial dissociation between core language processing areas on the one hand and areas involved in making inferences about others' mental states (ToM). We find no evidence of mentalizing in the core LH language network using a nonverbal ToM task, and we further find no selectivity for mentalizing over other kinds of social cognition. Linguistic analyses indicate that prior reports of overlap between the language and ToM networks may have been driven by confounds with linguistic variables independently known to drive language network activity. These results do not support a role for the language network in making inferences about others' mental states. The language "periphery"—consisting of the RH homotopic language areas and the language-responsive areas in the bilateral angular gyri—responds relatively more strongly than the core language network to conditions that encourage mentalizing. However, these stronger responses also extend to other kinds of social conditions and even nonsocial ones, which is consistent with these regions' role in social and even general visual-semantic processing.

Acknowledgements

We would like to thank Zach Mineroff, Melissa Kline, and Brianna Pritchett for helping with fMRI data collection; Rebecca Saxe and Ben Deen for comments on the earlier draft of the manuscript; and EvLab and TedLab members for helpful discussions. We would also like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT and its support team (Steve Shannon and Atsushi Takahashi).

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Funding

This work was supported by the National Institutes of Health (NIH) award R01-DC016607. EF was additionally supported by the NIH awards R01-DC016950 and U01-NS121471, by a grant from the Simons Foundation to the Simons Center for the Social Brain at the Massachusetts Institute of Technology, and by research funds from the McGovern Institute for Brain Research and the Department of Brain and Cognitive Sciences.

Conflict of interest statement: None declared.

Data availability

Data and code are available on OSF: https://osf.io/bzwm8/.

References

Apperly IA, Samson D, Carroll N, Hussain S, Humphreys G. Intact first-and second-order false belief reasoning in a patient with severely impaired grammar. *Soc Neurosci*. 2006:1(3–4):334–348.

Apperly IA, Samson D, Chiavarino C, Humphreys GW. Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. *J Cogn Neurosci*. 2004:16(10):1773–1784.

Ashburner J Karl JF. Unified segmentation. *Neuroimage*. 2005:26(3):839–851.

Astington JW, Baird JA. *Why language matters for theory of mind*. Oxford: Oxford University Press; 2005

Astington JW, Jenkins JM. A longitudinal study of the relation between language and theory-of-mind development. *Dev Psychol*. 1999:35(5):1311.

Bates D, Machler M, Bolker B and Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. 2015:67(1):1.48. https://doi.org/10.18637/jss.v067.i01.

Beeman MJ, Chiarello C. *Right hemisphere language comprehension: perspectives from cognitive neuroscience*. London: Psychology Press; 2013.

Bemis DK, Pylkkänen L. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cereb Cortex*. 2013:23(8):1859–1873.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001:29(4): 1165–1188.

Binder JR, Julie AF, Thomas AH, Robert WC, Stephen MR, and Thomas P. Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*. 1997:17(1):353–362.

Blank I, Kanwisher N, Fedorenko E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J Neurophysiol*. 2014:112(5): 1105–1118.

Blank I, Balewski Z, Mahowald K, Fedorenko E. Syntactic processing is distributed across the language system. *NeuroImage*. 2016a:127: 307–323.

Blank I, Duff MC, Brown-Schmidt S, Fedorenko E. Expanding the language network: domain-specific hippocampal recruitment during high-level linguistic processing. *BioRxiv*. 2016b:91900.

Bonner MF, Peelle JE, Cook PA, Grossman M. Heteromodal conceptual processing in the angular gyrus. *NeuroImage*. 2013:71:175–186.

Bottini G, Corcoran R, Sterzi R, Paulesu E, Schenone P, Scarpa P, Frackowiak RSJ, Frith D. The role of the right hemisphere in the interpretation of figurative aspects of language a positron emission tomography activation study. *Brain*. 1994:117(6): 1241–1253.

Braga RM, Buckner RL. Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron*. 2017:95(2):457–471.

Braga RM, DiNicola LM, Becker HC, Buckner RL. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J Neurophysiol*. 2020:124(5):1415–1448.

Branzi FM, Pobric G, Jung J, Lambon Ralph MA. The left angular gyrus is causally involved in context-dependent integration and associative encoding during narrative reading. *J Cogn Neurosci*. 2021:33(6):1082–1095.

Brothers L, Ring B. A neuroethological framework for the representation of minds. *J Cogn Neurosci*. 1992:4(2):107–118.

Bruneau EG, Dufour N, Saxe R. Social cognition in members of conflict groups: behavioural and neural responses in Arabs, Israelis and south Americans to each other's misfortunes. *Philo Trans R Soc B Biol Sci*. 2012a:367(1589):717–730.

Bruneau EG, Pluta A, Saxe R. Distinct roles of the 'shared pain'and 'theory of mind'networks in processing others' emotional suffering. *Neuropsychologia*. 2012b:50(2):219–231.

Bryan KL. Language prosody and the right hemisphere. *Aphasiology*. 1989:3(4):285–299.

Calvo N, Abrevaya S, Martínez Cuitiño M, Steeb B, Zamora D, Sedeño L, Ibáñez A, García AM. Rethinking the neural basis of prosody and non-literal language: spared pragmatics and cognitive compensation in a bilingual with extensive right-hemisphere damage. *Front Psychol*. 2019:10:570, 1–13. https://doi.org/10.3389/fpsyg.2019.00570.

Castelli F, Happé F, Frith U, Frith C. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*. 2000:12(3):314–325.

Castelli F, Frith C, Happé F, Frith U. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*. 2002:125(8):1839–1849.

Chai LR, Mattar MG, Blank IA, Fedorenko E, Bassett DS. Functional network dynamics of the language system. *Cereb Cortex*. 2016:26(11):4148–4159.

Champagne-Lavau M, Joanette Y. Pragmatics, theory of mind and executive functions after a right-hemisphere lesion: different patterns of deficits. *J Neurolinguistics*. 2009:22(5):413–426.

Davis CP, Yee E. Features, labels, space, and time: factors supporting taxonomic relationships in the anterior temporal lobe and thematic relationships in the angular gyrus. *Lang Cogn Neurosci*. 2019:34(10):1347–1357.

de Villiers JG, de Villiers PA. The role of language in theory of mind development. *Top Lang Disord*. 2014:34(4):313–328.

Deen B, Koldewyn K, Kanwisher N, Saxe R. Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb Cortex*. 2015:25(11):4596–4609.

Dennis M, Simic N, Bigler ED, Abildskov T, Agostino A, Taylor HG, Rubin K, Vannatta K, Gerhardt CA, Stancin T. Cognitive, affective, and conative theory of mind (ToM) in children with traumatic brain injury. *Dev Cogn Neurosci*. 2013:5:25–39.

Diehl JJ, Bennetto L, Young EC. Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *J Abnorm Child Psychol*. 2006:34(1):83–98.

DiNicola LM, Braga RM, Buckner RL. Parallel distributed networks dissociate episodic and social functions within the individual. *J Neurophysiol*. 2020:123(3):1144–1179.

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R. fMRI item analysis in a theory of mind task. *NeuroImage*. 2011:55(2):705–712.

Domínguez DJF, Nott Z, Horne K, Prangley T, Adams AG, Henry JD, Molenberghs P. Structural and functional brain correlates of theory of mind impairment post-stroke. *Cortex*. 2019:121: 427–442.

Dronkers NF, Ludy CA, Redfern BB. Pragmatics in the absence of verbal language: descriptions of a severe aphasic and a language-deprived adult. *J Neurolinguistics*. 1998:11(1–2):179–190.

Dufour N, Redcay E, Young L, Mavros PL, Moran JM, Triantafyllou C, Gabrieli JDE, Saxe R. Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLoS One*. 2013:8(9):e75468.

Farrer C, Frey SH, Van Horn JD, Tunik E, Turk D, Inati S, Grafton ST. The angular gyrus computes action awareness representations. *Cereb Cortex*. 2008:18(2):254–261.

Fedorenko E, Thompson-Schill SL. Reworking the language network. *Trends Cogn Sci*. 2014:18(3):120–126.

Fedorenko E, Hsieh P-J, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J Neurophysiol*. 2010:104(2):1177–1194.

Fedorenko E, Behr MK, Kanwisher N. Functional specificity for high-level linguistic processing in the human brain. *Proc Natl Acad Sci*. 2011:108(39):16428–16433.

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N. Neural correlate of the construction of sentence meaning. *Proc Natl Acad Sci*. 2016:113(41):E6256–E6262.

Fedorenko E, Blank I, Siegelman M, Mineroff Z. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*. 2020:203:104348.

Friston KJ, John A, Christopher DF, Poline J-B, John DH and Richard SJF. Spatial registration and normalization of images. *Human brain mapping*. 1995:3(3):165–189.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RSJ, Frith CD. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*. 1995:57(2): 109–128.

Friston KJ. Functional and effective connectivity: a review. *Brain Connectivity*. 2011:1(1):13–36.

Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith CD. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind'in verbal and nonverbal tasks. *Neuropsychologia*. 2000:38(1):11–21.

Gibson E. The dependency locality theory: a distance-based theory of linguistic complexity. In: Marantz A, Miyashita Y, O'Neil W, editors. *Image, language, brain*. Cambridge: MIT Press; 2000. pp. 95–106.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M. A multi-modal parcellation of human cerebral cortex. *Nature*. 2016:536(7615):171–178.

Graff, D., Kong, J., Chen, K., & Maeda, K. (2007). *English Gigaword third edition LDC2007T07*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2007T07

Grice HP. Logic and conversation. In: Cole P, Morgan JL, editors. *Syntax and semantics*, Vol. 3: Speech acts. New York: Academic Press; 1975. pp. 41–58.

Hale CM, Tager-Flusberg H. The influence of language on theory of mind: a training study. *Dev Sci*. 2003:6(3):346–359.

Hauptman M, Blank I, Fedorenko E. Non-literal language processing is jointly supported by the language and theory of mind networks: evidence from a novel meta-analytic fMRI approach. *BioRxiv*. 2022.

Heafield K, Pouzyrevsky I, Clark JH, Koehn P. Scalable modified Kneser-Ney language model estimation. In: *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2013. pp. 690–696.

Hein G, Singer T. I feel how you feel but not always: the empathic brain and its modulation. *Curr Opin Neurobiol*. 2008:18(2):153–158.

Humphreys GF, Ralph MAL, Simons JS. A unifying account of angular gyrus contributions to episodic and semantic cognition. *Trends Neurosci*. 2021:44(6):452–463.

Hutchison RM, Womelsdorf T, Allen EA, Bandettini PA, Calhoun VD, Corbetta M, Della Penna S, Duyn JH, Glover GH, Gonzalez-Castillo J. Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*. 2013:80:360–378.

Isik L, Kami K, David B and Nancy K. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*. 2017:114(43):E9145–E9152.

Ivanova A, Srikant S, Sueoka Y, Kean H, Dhamala R, O'Reill U-M, Bers MU and Fedorenko E. Comprehension of computer code relies primarily on domain-general executive resources. *eLife*. 2020:9:e58906.

Ivanova AA, Siegelman M, Cheung C, Pongos ALA, Kean H, Fedorenko E. The language network responds robustly to sentences regardless of task. in prep.

Ivanova A. The role of language in broader human cognition: Evidence from neuroscience. PhD Thesis, MIT. in prep b.

Jacoby N, Bruneau E, Koster-Hale J, Saxe R. Localizing pain matrix and theory of mind networks with both verbal and non-verbal stimuli. *NeuroImage*. 2016:126:39–48.

Joseph R. The right cerebral hemisphere: emotion, music, visual-spatial skills, body-image, dreams, and awareness. *J Clin Psychol*. 1988:44(5):630–673.

Julian JB, Evelina F, Jason W and Nancy K. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*. 2012:60(4):2357–2364.

Just MA, Carpenter PA. A theory of reading: from eye fixations to comprehension. *Psychol Rev*. 1980:87(4):329–354.

Kaakinen JK, Olkoniemi H, Kinnari T, Hyönä J. Processing of written irony: an eye movement study. *Discourse Process*. 2014:51(4):287–311.

Kamps FS, Richardson H, Murty NAR, Kanwisher N, Saxe R. Using child-friendly movie stimuli to study the development of face, place, and object regions from age 3 to 12 years. *Hum Brain Mapp*. 2022:43(9):2782–2800.

Kaplan JA, Brownell HH, Jacobs JR, Gardner H. The effects of right hemisphere damage on the pragmatic interpretation of conversational remarks. *Brain Lang*. 1990:38(2):315–333.

Koster-Hale J, Saxe R. Theory of mind: a neural prediction problem. *Neuron*. 2013:79(5):836–848.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*. 2009:12(5):535–540.

LeBel A, Jain S, Huth AG. Voxelwise encoding models show that cerebellar language representations are highly conceptual. *J Neurosci*. 2021:41(50):10341–10355.

Lee SS, Dapretto M. Metaphorical vs. literal word meanings: fMRI evidence against a selective role of the right hemisphere. *NeuroImage*. 2006:29(2):536–544.

Lindell AK. In your right mind: right hemisphere contributions to language processing and production. *Neuropsychol Rev*. 2006:16(3):131–148.

Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, Kean H, Jouravlev O, Rakocevic L, Pritchett B, Siegelman M, et al. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*. 2022:9(529).

Lombardo MV, Chakrabarti B, Bullmore ET, Baron-Cohen S, Consortium MRCA. Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *NeuroImage*. 2011:56(3):1832–1838.

Lopopolo A, Frank SL, den Bosch A, Willems RM. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS One*. 2017:12(5):e0177794.

Mahowald K and Fedorenko E. Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *Neuroimage*. 2016:139:74.93.

Malik-Moraleda S, Ayyash D, Gallée J, Affourtit J, Hoffman M, Mineroff Z, Jouravlev O, Fedorenko E. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*. 2022:25(8):1014–1019.

Mar RA. The neural bases of social cognition and story comprehension. *Annu Rev Psychol*. 2011:62:103–134.

Marcus MP, Santorini B, Marcinkiewicz MA. Building a large annotated corpus of English: the Penn treebank. *Comput Linguist*. 1993:19(2):313–330.

Martin KC, Seydell-Greenwald A, Berl MM, Gaillard WD, Turkeltaub PE, Newport EL. A weak shadow of early life language processing persists in the right hemisphere of the mature brain. *Neurobiol Lang*. 2022:3(3):364–385.

Martín-Rodríguez JF, León-Carrión J. Theory of mind deficits in patients with acquired brain injury: a quantitative review. *Neuropsychologia*. 2010:48(5):1181–1191.

Mason RA, Just MA. Differentiable cortical networks for inferences concerning people's intentions versus physical causality. *Hum Brain Mapp*. 2011:32(2):313–329.

Matchin W, Liao C-H, Gaston P, Lau E. Same words, different structures: an fMRI investigation of argument relations and the angular gyrus. *Neuropsychologia*. 2019:125:116–128.

Miller CA. Developmental relationships between language and theory of mind. *American Journal of Speech-Langu age Pathology*. 2006:15(2):142–154.

Mitchell RLC, Crow TJ. Right hemisphere language functions and schizophrenia: the forgotten hemisphere? *Brain*. 2005:128(5):963–978.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & others. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc Natl Acad Sci*, 114(18), E3669–E3678.

Nguyen L, van Schijndel M, Schuler W. Accurate unbounded dependency recovery using generalized categorial grammars. In: *Proceedings of COLING 2012*. Association for Computational Linguistics; 2012.

Nieto-Castanon A. *Handbook of functional connectivity Magnetic Resonance Imaging methods in CONN*. Hilbert Press, 2020.

Nieto-Castañón A, Fedorenko E. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*. 2012:63(3):1646–1669.

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. 1971:9(1):97–113.

Pallier C, Devauchelle A-D, Dehaene S. Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci*. 2011:108(6): 2522–2527.

Paunov A, Blank I, Fedorenko E. Functionally distinct language and theory of mind networks are synchronized at rest and during language comprehension. *J Neurophysiol*. 2019:121(4):1244–1265.

Paunov A, Blank IA, Jouravlev O, Mineroff Z, Gallée J, Fedorenko E. Differential tracking of linguistic vs. mental state content in naturalistic stimuli by language and theory of mind (ToM) brain networks. *Neurobiol Lang*. 2022:3(3):413–440.

Peterson CC, Siegal M. Insights into theory of mind from deafness and autism. *Mind Lang*. 2000:15(1):123–145.

Price AR, Bonner MF, Peelle JE, Grossman M. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *J Neurosci*. 2015:35(7):3276–3284.

Pritchett BL, Hoeflin C, Koldewyn K, Dechter E, Fedorenko E. High-level language processing regions are not engaged in action observation or imitation. *J Neurophysiol*. 2018:120(5):2555–2570.

Quillen IA, Yen M, Wilson SM. Distinct neural correlates of linguistic and non-linguistic demand. *Neurobiol Lang*. 2021:2(2):202–225.

Rajimehr R, Firoozi A, Rafipoor H, Abbasi N, Duncan J. Complementary hemispheric lateralization of language and social processing in the human brain. *Cell Reports*. 2022:41(6):111617.

Rapp AM, Leube DT, Erb M, Grodd W, Kircher TTJ. Laterality in metaphor processing: lack of evidence from functional magnetic resonance imaging for the right hemisphere theory. *Brain Lang*. 2007:100(2):142–149.

Rapp AM, Mutschler DE, Erb M. Where in the brain is nonliteral language? A coordinate-based meta-analysis of functional magnetic resonance imaging studies. *NeuroImage*. 2012:63(1):600–610.

Rayner K, Kambe G, Duffy SA. The effect of clause wrap-up on eye movements during reading. *Q J Exp Psychol A*. 2000:53(4): 1061–1080.

Regel S, Gunter TC, Friederici AD. Isn't it ironic? An electrophysiological exploration of figurative language processing. *J Cogn Neurosci*. 2011:23(2):277–293.

Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. Development of the social brain from age three to twelve years. *Nat Commun*. 2018:9(1):1–12.

Richardson H, Koster-Hale J, Caselli N, Magid R, Benedict R, Olson H, Pyers J, Saxe R. Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *Nat Commun*. 2020:11(1):1–13.

Roberts C. Information structure: towards an integrated formal theory of pragmatics. *Semant Pragmat*. 2012:5:1–6.

Ross ED, Mesulam M-M. Dominant language functions of the right hemisphere? Prosody and emotional gesturing. *Arch Neurol*. 1979:36(3):144–148.

Ruby P, Decety J. What you believe versus what you think they believe: a neuroimaging study of conceptual perspective-taking. *Eur J Neurosci*. 2003:17(11):2475–2480.

Ruffman T, Slade L, Rowlandson K, Rumsey C, Garnham A. How language relates to belief, desire, and emotion understanding. *Cogn Dev*. 2003:18(2):139–158.

Saxe R. Uniquely human social cognition. *Curr Opin Neurobiol*. 2006:16(2):235–239.

Saxe R. The right temporo-parietal junction: a specific brain region for thinking about thoughts. In: *Handbook of theory of mind*; 2010. pp. 1–35.

Saxe R, Kanwisher N. People thinking about thinking people: the role of the temporo-parietal junction in "Theory of Mind". *NeuroImage*. 2003:19(4):1835–1842.

Saxe R, Powell LJ. It's the thought that counts: specific brain regions for one component of theory of mind. *Psychol Sci*. 2006:17(8): 692–699.

Saxe R, Wexler A. Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*. 2005:43(10): 1391–1399.

Saxe R, Carey S, Kanwisher N. Understanding other minds. *Annu Rev Psychol*. 2004:55:87–124.

Saxe R, Brett M, Kanwisher N. Divide and conquer: a defense of functional localizers. *NeuroImage*. 2006:30(4):1088–1096.

Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown EN, Saxe R. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One*. 2009:4(3):e4869.

Schurz M, Radua J, Aichhorn M, Richlan F, Perner J. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev*. 2014:42:9–34.

Schurz M, Tholen MG, Perner J, Mars RB, Sallet J. Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: a review using probabilistic atlases from different imaging modalities. *Hum Brain Mapp*. 2017:38(9): 4788–4805.

Schuster S, Hawelka S, Hutzler F, Kronbichler M, Richlan F. Words in context: the effects of length, frequency, and predictability on brain responses during natural reading. *Cereb Cortex*. 2016:26(10): 3889–3904.

Scott TL, Gallée J, Fedorenko E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn Neurosci*. 2017:8(3):167–176.

Seghier ML. The angular gyrus: multiple functions and multiple subdivisions. *Neuroscientist*. 2013:19(1):43–61.

Shain C. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In: *Proceedings of the 2019 conference of the north American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*. Association for Computational Linguistics; 2019. pp. 4086–4094

Shain C, Blank I, van Schijndel M, Schuler W, Fedorenko E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*. 2020:138: 107307.

Shain C, Blank IA, Fedorenko E, Gibson E, Schuler W. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *J Neurosci*. 2022:42(39): 7412–7430.

Shamay-Tsoory SG, Harari H, Aharon-Peretz J, Levkovitz Y. The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*. 2010:46(5):668–677.

Shibata M, Toyomura A, Itoh H, Abe J. Neural substrates of irony comprehension: a functional MRI study. *Brain Res*. 2010:1308: 114–123.

Singer T, Lamm C. The social neuroscience of empathy. *Ann N Y Acad Sci*. 2009:1156(1):81–96.

Slade L, Ruffman T. How language does (and does not) relate to theory of mind: a longitudinal study of syntax, semantics, working memory and false belief. *Br J Dev Psychol*. 2005:23(1): 117–141.

Small H, Lipkin B, Affourtit J, Pongos A, Fedorenko E. Differential selectivity of the left and right hemisphere language regions for non-linguistic processing. In: *Proceedings of the thirteenth annual meeting of the Society for the Neurobiology of language*. The Society for the Neurobiology of Language; 2021. p. 264.

Sommer M, Döhnel K, Sodian B, Meinhardt J, Thoermer C, Hajak G. Neural correlates of true and false belief reasoning. *NeuroImage*. 2007:35(3):1378–1384.

Sperber D, Wilson D. Précis of relevance: communication and cognition. *Behav Brain Sci*. 1987:10(4):697–710.

Sprong M, Schothorst P, Vos E, Hox J, Van Engeland H. Theory of mind in schizophrenia: meta-analysis. *Br J Psychiatry*. 2007:191(1): 5–13.

Tager-Flusberg H, Paul R, Lord C. Language and communication in autism. In: *Handbook of autism and pervasive developmental disorders*. Hoboken: John Wiley & Sons. Vol. 1; 2005. pp. 335–364.

Thothathiri M, Kimberg DY, Schwartz MF. The neural basis of reversible sentence comprehension: evidence from voxel-based lesion symptom mapping in aphasia. *J Cogn Neurosci*. 2012:24(1): 212–222.

Uddin LQ, Supekar K, Amin H, Rykhlevskaia E, Nguyen DA, Greicius MD, Menon V. Dissociable connectivity within human angular gyrus and intraparietal sulcus: evidence from functional and structural connectivity. *Cereb Cortex*. 2010:20(11):2636–2646.

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, Wu-Minn H, C P Consortium. The WU-Minn human connectome project: an overview. *NeuroImage*. 2013:80:62–79.

Van Lancker D. Rags to riches: our increasing appreciation of cognitive and communicative abilities of the human right cerebral hemisphere. *Brain Lang*. 1997:57(1):1–11.

Van Overwalle F. Social cognition and the brain: a meta-analysis. *Hum Brain Mapp*. 2009:30(3):829–858.

van Schijndel M, Exley A, Schuler W. A model of language processing as hierarchic sequential prediction. *Top Cogn Sci*. 2013:5(3): 522–540.

Varley R, Siegal M, Want SC. Severe impairment in grammar does not preclude theory of mind. *Neurocase*. 2001:7(6):489–493.

Vogeley K, Bussfeld P, Newen A, Herrmann S, Happé F, Falkai P, Maier W, Shah NJ, Fink GR, Zilles K. Mind reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage*. 2001:14(1): 170–181.

Wellman HM, Cross D, Watson J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev*. 2001:72(3): 655–684.

Willems RM, Benn Y, Hagoort P, Toni I, Varley R. Communicating without a functioning language system: implications for the role of language in mentalizing. *Neuropsychologia*. 2011:49(11): 3130–3135.

Willems RM, der Haegen L, Fisher SE, Francks C. On the other hand: including left-handers in cognitive neuroscience and neurogenetics. *Nat Rev Neurosci*. 2014:15(3):193.

Wimmer H, Perner J. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*. 1983:13(1):103–128.

Winner E, Brownell H, Happé F, Blum A, Pincus D. Distinguishing lies from jokes: theory of mind deficits and discourse interpretation in right hemisphere brain-damaged patients. *Brain Lang*. 1998:62(1):89–106.

Young L, Dodell-Feder D, Saxe R. What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*. 2010:48(9):2658–2664.

Zaidel DW. Hemispheric asymmetry in long-term semantic relationships. *Cogn Neuropsychol*. 1987:4(3):321–332.