Title

Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks.

Journal

Neurobiology of Language, 3(3)

Authors

Paunov, Alexander
Blank, Idan
Jouravlev, Olessia
et al.

Publication Date

2022

DOI

10.1162/nol_a_00071

Peer reviewed

Neurobiology of Language

an open access journal

RESEARCH ARTICLE

Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks

Alexander M. Paunov[1,2], Idan A. Blank[1,3], Olessia Jouravlev[1,4,5], Zachary Mineroff[1,4,6], Jeanne Gallée[7], and Evelina Fedorenko[1,4,7]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, USA
[2]Cognitive Neuroimaging Unit, INSERM, CEA, CNRS, Université Paris-Saclay, NeuroSpin Center, 91191Gif/Yvette, France
[3]Department of Psychology, UCLA, Los Angeles, CA, USA
[4]McGovern Institute for Brain Research, MIT, Cambridge, MA, USA
[5]Institute for Cognitive Science, Carleton University, Ottawa, ON, Canada
[6]Eberly Center for Teaching Excellence & Educational Innovation, Carnegie Mellon University, Pittsburgh, PA, USA
[7]Program in Speech and Hearing Bioscience and Technology, Harvard University, Boston, MA, USA

Keywords: language, theory of mind, social cognition, inter-subject correlations, naturalistic fMRI

ABSTRACT

Language and social cognition, especially the ability to reason about mental states, known as *theory of mind* (ToM), are deeply related in development and everyday use. However, whether these cognitive faculties rely on distinct, overlapping, or the same mechanisms remains debated. Some evidence suggests that, by adulthood, language and ToM draw on largely distinct—though plausibly interacting—cortical networks. However, the broad topography of these networks is similar, and some have emphasized the importance of social content / communicative intent in the linguistic signal for eliciting responses in the language areas. Here, we combine the power of individual-subject functional localization with the naturalistic-cognition inter-subject correlation approach to illuminate the language–ToM relationship. Using functional magnetic resonance imaging (fMRI), we recorded neural activity as participants (n = 43) listened to stories and dialogues with mental state content (+linguistic, +ToM), viewed silent animations and live action films with mental state content but no language (−linguistic, +ToM), or listened to an expository text (+linguistic, −ToM). The ToM network robustly tracked stimuli rich in mental state information regardless of whether mental states were conveyed linguistically or non-linguistically, while tracking a +linguistic / −ToM stimulus only weakly. In contrast, the language network tracked linguistic stimuli more strongly than (a) non-linguistic stimuli, and than (b) the ToM network, and showed reliable tracking even for the linguistic condition devoid of mental state content. These findings suggest that in spite of their indisputably close links, language and ToM dissociate robustly in their neural substrates—and thus plausibly cognitive mechanisms—including during the processing of rich naturalistic materials.

INTRODUCTION

Language and social cognition, especially the ability to reason about mental states, known as *theory of mind* (ToM), are deeply related in human development, everyday use, and possibly evolution. After all, language use is a communicative behavior, which is a kind of cooperative

Citation: Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (2022). Differential tracking of linguistic vs. mental state content in naturalistic stimuli by language and theory of mind (ToM) brain networks. *Neurobiology of Language*, *3*(3), 413–440. https://doi.org/10.1162/nol_a_00071

DOI: https://doi.org/10.1162/nol_a_00071

Supporting Information: https://doi.org/10.1162/nol_a_00071

Received: 26 April 2021
Accepted: 11 April 2022

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Alexander M. Paunov
alexander.paunov@gmail.com

Handling Editor:
Roel Willems

The MIT Press

Theory of mind (ToM): The social cognitive ability to reason about others' minds in terms of unobservable mental states like belief, desire, and intention.

behavior, and cooperative behaviors are, in turn, a kind of social behavior (e.g., Grice, 1968, 1975; Sperber & Wilson, 1986). Construed this way, language can hardly be encapsulated from social cognition (cf. Fodor, 1983); it is subsumed *within* social cognition. Interpreting linguistic signals bears key parallels to the interpretation of other intentional behaviors (e.g., Grice, 1975; Sperber & Wilson, 1986). Communicative utterances, like other behaviors, are assumed to have goals, and conversation partners are assumed to pursue these goals rationally. Furthermore, everyday discourse appears to be dominated by information about other people (e.g., Dunbar et al., 1997), and the need to keep track of others' social record has been proposed as a key driver of language evolution (e.g., Dunbar, 2004; Nowak & Highfield, 2011; Nowak & Sigmund, 2005; Sommerfeld et al., 2008). Lastly, evidence of others' mental states conveyed through language is arguably richer and certainly more direct / less ambiguous than what can be inferred from non-linguistic intentional behavior alone: Trying to infer the beliefs guiding someone's actions can be obviated by their telling you what those beliefs are.

In spite of this deep relationship, language processing and social cognitive processing appear to draw on distinct neural mechanisms. Language processing engages a network of left-lateralized brain regions in lateral frontal and temporal cortex. These regions support lexico-semantic processing (word meanings) and combinatorial morphosyntactic and semantic processing (e.g., Anderson et al., 2019; Bautista & Wilson, 2016; Blank et al., 2016; Fedorenko et al., 2010, 2012, 2016, 2020; Reddy & Wehbe, 2020). In contrast to their robust and consistent responses to linguistic stimuli, these regions do not respond to a wide range of non-linguistic cognitive processes, including arithmetic processing, music perception, executive function tasks, the processing of computer code (e.g., Amalric & Dehaene, 2018; Fedorenko et al., 2011; Ivanova, Srikant, et al., 2020; Liu et al., 2020; Monti et al., 2012), and—critically—perceptual and cognitive social stimuli and tasks (e.g., Jouravlev et al., 2019; Paunov, 2019; Paunov et al., 2022; Pritchett et al., 2018; see Fedorenko & Varley, 2016, for a review).

On the other hand, attribution of mental states, such as beliefs, desires, and intentions, engages a network of brain regions in bilateral temporoparietal cortex and anterior and posterior regions along the cortical midline. These responses generalize across the type of mental state, its specific content or format (linguistic vs. pictorial), and the source of evidence for it (e.g., Castelli et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000; Jacoby et al., 2016; Kandylaki et al., 2015; Ruby & Decety, 2003; Saxe & Kanwisher, 2003; Saxe, Schulz, & Jiang, 2006; Vogeley et al., 2001; see Koster-Hale & Saxe, 2013, for review). By adulthood, these regions, and especially the most selective component of the ToM network, the right temporoparietal junction (RTPJ), do not respond to social stimuli, like faces, voices, or biological motion (e.g., Deen et al., 2015), to general executive demands (e.g., Saxe, Schulz, & Jiang, 2006), to physical or broadly social attributes of agents, or to attribution of bodily sensations of pain and hunger (e.g., Bruneau et al., 2012; Jacoby et al., 2016; Saxe & Powell, 2006).

Investigations of developmental and acquired disorders have provided convergent support for the dissociability of language and ToM mechanisms. Some individuals with even severe aphasia appear to retain the capacity for mental state reasoning as long as nonverbal materials are used (e.g., Apperly et al., 2006; Varley & Siegal, 2000; Varley et al., 2001; Willems et al., 2011; see Fedorenko & Varley, 2016, for review). And at least some individuals with social, ToM-related impairments (e.g., some individuals with autism spectrum disorders) show preservation of lexical and syntactic linguistic abilities (e.g., Frith & Happé, 1994; Tager-Flusberg et al., 2005; Wilkinson, 1998).

Yet, given the deep relationship between language and ToM, ToM mechanisms must be engaged during language processing at least sometimes. When does this happen? The role

Pragmatics:
The context-based inference of intended (nonliteral) meaning.

of social-cognitive mechanisms in language comprehension—and thus the degree of segregation between social and linguistic mechanisms—can be examined with respect to two aspects of the linguistic signal. One concerns the role of mental state inference in language comprehension generally, whether or not the message content is about mental states. This question is at the core of the field of pragmatics, which aims to understand how communicative intent—a form of ToM inference—guides linguistic interpretation (e.g., Grice, 1957/1991, 1968, 1975). And the other relates to the use of language to express information about the mental states of agents, either directly or through descriptions of physical events, which prompt mental state attribution (e.g., Fletcher et al., 1995; Gallagher et al., 2000; Jacoby et al., 2016; Saxe & Kanwisher, 2003; Saxe, Schulz, & Jiang, 2006). In Figure 1, we schematically illustrate these different kinds of social-cognitive demands—understanding the *communicative intent* of the person generating the linguistic output and the processing of the *linguistic content*—across three common contexts for language processing.

Past work in cognitive neuroscience has investigated both of these ToM demands in language processing. Some studies have manipulated the difficulty of inferring the communicative intent of a speaker by examining paradigmatic cases of nonliteral language, from irony (e.g., Spotorno et al., 2012), to indirect speech (e.g., Feng et al., 2017; van Ackeren et al., 2012) and other forms of conversational implicature (e.g., Feng et al., 2021; Jang et al., 2013; see Hagoort & Levinson, 2014, for review). These studies have reported stronger responses in the ToM network for the critical, nonliteral stimuli compared to literal controls. However, delimiting the scope of pragmatic inference is a long-standing challenge, raising the question of whether it is possible to draw a boundary between decoded (literal) and inferred meaning, that is, between semantics and pragmatics (e.g., Jackendoff, 2009). The lack of clearly defined boundaries for the construct of pragmatics poses empirical challenges. For one, pragmatic inference need not be limited to linguistic communication: It can equally be

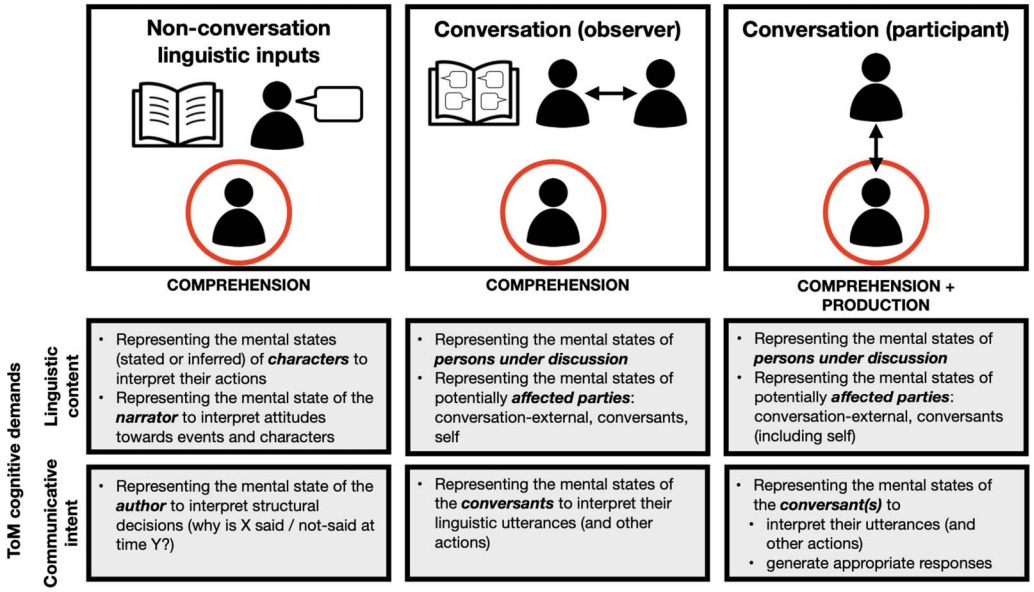

Figure 1. Cognitive demands on ToM processing across three different contexts of discourse-level linguistic processing (reading or listening to narratives or expository texts, reading or listening to a conversation, and directly participating in face-to-face conversation) with respect to understanding the communicative intent of the person generating the linguistic output, and the linguistic content of the materials. The comprehender is circled in red. Whereas specific demands differ across contexts, any form of language use arguably involves attribution of communicative intent (pragmatics) in the service of comprehension and/or production.

present in other forms of cooperative information transfer between individuals, such as the interpretation of communicative gestures or, more relevant in the present context, understanding non-linguistic "stories" such as nonverbal animated or live-action films. At the same time, within the narrower context of linguistic communication, it is implausible that all forms of context-based inference of meaning recruit mental state reasoning (cf. Sperber & Wilson, 2002). Phenomena that require context-based inferences include not just the paradigmatic instances of nonliteral meaning, such as irony, indirect requests, hyperbole, and other *conversational implicatures* (Grice, 1975)—which do plausibly require ToM (at least often, if not necessarily)—but also relatively "low-level" and ubiquitous phenomena such as pronoun resolution, or lexical and syntactic disambiguation. Indeed, functional magnetic resonance imaging (fMRI) studies that have focused on the latter kinds of phenomena do not report ToM network engagement (e.g., Bahlmann et al., 2007; Hammer et al., 2011; January et al., 2009; Klepousniotou et al., 2014; Mason et al., 2003; McMillan et al., 2012; Rodd et al., 2005; Snijders et al., 2009).

Other studies have varied the amount of mental state content in verbal vignettes or stories (e.g., Ferstl & von Cramon, 2002; Saxe & Powell, 2006; for relevant behavioral evidence, see Bischetti et al., 2019; Lecce et al., 2019) and reported stronger activity in the ToM network for linguistic materials rich in mental state content. So, both kinds of demands appear capable of recruiting mental state reasoning, but delineating the conditions under which ToM mechanisms get engaged during linguistic processing remains an area of active investigation.

In the current study, we take a step back and re-examine the separability of linguistic and ToM mechanisms. Many studies to date have deliberately aimed to isolate the language or the ToM system rather than probe their relationship (cf. Braga et al., 2020; Deen et al., 2015; Paunov et al., 2019) and have relied on traditional experimentally controlled paradigms. As a result, we know that language and social cognition *can* dissociate, under appropriate experimental conditions. But do they, in fact, dissociate in everyday, naturalistic cognition? Paunov et al. (2019) recently used fMRI to examine inter-regional functional correlations within and between the language and ToM networks during naturalistic cognition paradigms, like story comprehension, and found that in spite of their dissociability (stronger within- than between-network correlations), the language and ToM networks also showed a significant amount of synchronization in their neural activity. These results point to some degree of functional integration between the networks. Furthermore, several prior studies have argued for the importance of social content / communicative intent in the linguistic signal for eliciting responses in the language areas (e.g., Mellem et al., 2016; Redcay et al., 2016). In line with current emphasis in the field on the importance of going beyond carefully controlled experimental materials in the testing of hypotheses about human cognitive architecture (e.g., Blank et al., 2014; Hasson et al., 2018; Huth et al., 2016), we use rich naturalistic materials. In contrast to Paunov et al. (2019), we rely on the inter-subject correlation (ISC) approach (e.g., Hasson et al., 2004, 2008), which allows us to not only examine the degree of dissociability and interaction between the networks, but also ask what aspects of the stimulus each network responds to by varying the format and content of the materials.

Inter-subject correlation:
A model-free method for estimating the degree of "tracking" of a (usually complex, naturalistic) stimulus across individuals.

MATERIALS AND METHODS

General Approach

Following Blank and Fedorenko (2017, 2020), we combine two powerful methodologies that have previously been productively applied separately in the domains of language and social cognition. In particular, we use *functional localization* (e.g., Brett et al., 2002; Fedorenko et al.,

2010; Saxe, Brett, & Kanwisher, 2006) to identify the two networks of interest in individual subjects, and *inter-subject correlations* (e.g., Hasson et al., 2004, 2008) to examine the degree to which these networks "track" different stimulus features during the processing of rich naturalistic materials. Here, we highlight key strengths of each approach and consider their synergistic advantages in the context of our research goals.

Naturalistic paradigms have become a crucial complement to traditional, task-based studies in cognitive neuroscience. The obvious advantage of naturalistic paradigms is their high ecological validity: By giving up a measure of experimental control, it becomes possible to study cognition "in the wild" (Blank & Fedorenko, 2017; for general discussions, see, e.g., Nastase et al., 2020; Sonkusare et al., 2019). In particular, one can examine how coherent and structured mental representations are extracted from rich and noisy perceptual inputs, which is what happens in everyday cognition. This is in contrast to artificially isolating various features of these perceptual inputs, as is typically done in constrained experimental tasks. Naturalistic paradigms have been argued to elicit more reliable responses compared to traditional, task-based paradigms (Hasson et al., 2010), perhaps because they are generally more engaging, and can enable discoveries of functional relationships among brain regions and networks that are altogether missed in more constrained settings (e.g., Gallivan et al., 2009). Another advantage of naturalistic paradigms is their hypothesis-free nature. Through the use of naturalistic materials, researchers impose minimal design constraints to investigate the domain of interest in a manner that is maximally unbiased by prior theoretical assumptions. In effect, they are letting the data speak for itself.

However, naturalistic paradigms also come with an inherent analytic challenge: How do we make sense of data acquired without the typical constraints of standard hypothesis-driven modeling approaches? Hasson et al. (2004) pioneered an approach to tackle this challenge, known as the *ISC approach* (see Hasson et al., 2008, for an overview), which we adopt in the current study. The key insight behind the ISC approach is that we can model any given participant's fMRI signal time series using another participant's or other participants' time series: If a voxel, brain region, or brain network "tracks" features of the stimulus during which a time series is obtained, then fMRI signal fluctuations will be stimulus-locked, resulting in similar time courses across participants (i.e., high inter-subject correlations).

ISCs have been used in several studies of narrative comprehension (e.g., Blank & Fedorenko, 2017, 2020; Honey et al., 2012; Lerner et al., 2011; Regev et al., 2013; Schmälzle et al., 2015; Silbert et al., 2014; Wilson et al., 2007), and whole-brain voxel-wise analyses have revealed high ISCs across large swaths of cortex that resemble the union of the language and ToM networks. On their own, these results might be taken as prima facie evidence for non-dissociability of language and ToM, given that the two networks appear to be jointly recruited. And insofar as the mental processes recruited in narrative comprehension recapitulate those used in everyday abstract cognition—an assumption that, we take it, partially justifies the interest in narratives in cognitive science and neuroscience (e.g., Finlayson & Winston, 2011; Willems et al., 2020)—the results may be taken to suggest the non-dissociability of language and ToM more generally.

However, it is difficult to draw inferences from these studies about the *relative* contributions of the language and ToM networks to narrative comprehension for two reasons. First, in whole-brain analyses, ISCs are computed on a voxel-wise basis: Individual brains are normalized to a stereotaxic template, and one-to-one voxel correspondence across individuals is then assumed in computing the ISCs. This approach is problematic because (i) inter-individual variability is

well established in the high-level association cortex (e.g., Fischl et al., 2008; Frost & Goebel, 2012; Tahmasebi et al., 2012), so any given voxel may belong to functionally distinct regions across participants; and (ii) there is no independent criterion based on which an anatomical location can be interpreted as belonging to the language vs. the ToM network, thus necessitating reliance on the fallacious "reverse inference" (Poldrack, 2006) to interpret the resulting topography (see Fedorenko, 2021, for discussion). And second, traditional whole-brain analyses typically include all voxels that showed significant (above baseline) ISCs, thus potentially obscuring large differences in effect sizes (cf. Blank & Fedorenko, 2017; see Chen et al., 2017, for a general discussion of the importance of considering effect sizes in interpreting fMRI findings). Combining the ISC approach with individual-participant functional localization enables us to identify and directly compare the networks of interest (including with respect to effect sizes), as well as to relate the findings straightforwardly to the prior literature on the language and ToM networks. In the current study, we therefore identified the language and ToM networks using well-established functional localizers (Fedorenko et al., 2010; Saxe & Kanwisher, 2003), and then examined the degree of inter-subject synchronization in those regions during the processing of diverse naturalistic linguistic and non-linguistic conditions varying in the presence of mental state content. If the language and ToM networks are dissociable during naturalistic cognition, we would expect the language regions to track linguistic stimuli, including those that lack mental state content, and the ToM regions to track stimuli that have mental state content, including both linguistic and non-linguistic ones.

Functional localizer: A task used to identify specialized regions of interest (fROIs) within individuals to establish functional correspondence of regions and networks across individuals.

Overall Experimental Design and Statistical Analyses

Our overall design and analytic strategy were as follows: Participant-specific regions that responded more strongly during the reading of sentences compared with lists of nonwords were defined as regions of interest comprising the language network. Similarly, regions that responded more strongly to stories about others' beliefs vs. stories about physical reality were defined as regions of interest comprising the ToM network (see the Stimuli and Procedure section for details). Whereas the precise anatomical locations of these regions were allowed to vary across participants, their overall topography was constrained by independently derived criteria to establish functional correspondence across brain regions of different participants (e.g., Fedorenko et al., 2010; Julian et al., 2012).

Activity in these two sets of brain regions was recorded with fMRI while participants listened to or watched a series of naturalistic stimuli, as detailed below. For each region in each network, our critical dependent variable was the strength of the correlation between each participant's time series and the average time series from the rest of the participants. The group-averaged ISC in each region was tested for significance via a permutation test of the time series data. For our critical analysis, all individual ISC values were modeled using a linear mixed-effects (LME) regression with participant, brain region, and stimulus (what we call "condition" below) as random effects.

Participants

Forty-seven native English speakers (age 19–48, $M = 24.5$, $SD = 5.08$; 30 female) from MIT and the surrounding Boston community participated for payment. Forty participants were right-handed, as determined by the Edinburgh handedness inventory (Oldfield, 1971) or by self-report ($n = 1$). All seven left-handed participants showed typical left lateralization in the language localizer task described below (see Willems et al., 2014, for arguments to include left handers in cognitive neuroscience research). Four participants were excluded from the

analyses due to poor quality of the localizer data (2 for ToM localizer, 1 for language localizer, and 1 for both), with the exclusion criterion defined as fewer than 100 suprathreshold voxels (at the $p < 0.001$ uncorrected whole-brain threshold) across the respective network's masks (see below), bringing the number of participants included in the critical analyses to 43. All participants gave informed written consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects (COUHES).

Stimuli and Procedure

Each participant completed a language localizer, a ToM localizer, a localizer for the domain-general MD system (used in a replication analysis, as described below), and a subset, or all, of the critical naturalistic stimuli ("conditions") (between 1 and 7) due to scan duration constraints (18 participants completed all 7 conditions of interest, 1 participant completed 6 conditions, 10 participants completed 5 conditions, 1 participant completed 4 conditions, 1 participant completed 3 conditions, 7 participants completed 2 conditions, and 5 participants completed 1 condition). Each condition was presented to between 28 and 32 participants (see Table 1). Each stimulus, lasting ~5–7 min (see Table 1 for precise durations), was preceded and followed by fixation (16 s and 32 s, respectively). Finally, 10 participants performed a resting state scan, used in one of the reality-check analyses, as described below. For the language localizer, 36/43 participants completed it in the same session as the critical conditions, the remaining 7 participants completed it in an earlier session. Similarly, for the ToM and MD localizers, 37/43 participants completed them in the same session as the critical conditions, the remaining 6 participants completed them in an earlier session. We will now describe the localizers and the critical experiment in more detail.

Table 1. Naturalistic conditions in each of the three condition types of interest and a reality-check condition (resting state), including durations, and number of participants per condition. Durations include 16 s fixations at the beginning and 32 s end of the scan (48 s total).

Condition	Duration	*N*
+Lang +ToM		
Story	5 m 16 s	31
Audio play	6 m 14 s	29
Dialogue	5 m 35 s	28
−Lang +ToM		
Animated short film	5 m 48 s	32
Live action movie clip	6 m 10 s	30
Intentional shapes animation	4 m 50 s	30
+Lang −ToM		
Expository text	7 m 6 s	28
−Lang −ToM (control)		
Resting state	5 m 0 s	10

Language localizer task

The task used to localize the language network is described in detail in Fedorenko et al. (2010) and targets brain regions that support high-level language processing. Briefly, we used a reading task contrasting sentences and lists of unconnected, pronounceable nonwords (Figure 2) in a blocked design with a counterbalanced condition order across runs. Stimuli were presented one word / nonword at a time. Participants read the materials passively (we included a button-pressing task at the end of each trial, to help participants remain alert). As discussed in the introduction, this localizer is robust to task manipulations (e.g., Fedorenko et al., 2010; Ivanova, Siegelman, et al., 2020; Scott et al., 2017). Moreover, this localizer identifies the same regions that are localized with a broader contrast, between listening to natural speech and its acoustically degraded version (Ayyash et al., 2022; Scott et al., 2017). All participants completed two runs, each lasting 358 s and consisting of 8 blocks per condition and 5 fixation blocks. (A version of this localizer is available from https://evlab.mit.edu/funcloc/.)

ToM localizer task

The task used to localize the ToM network is described in detail in Saxe and Kanwisher (2003) and targets brain regions that support reasoning about others' mental states. Briefly, the task was based on the classic false belief paradigm (Wimmer & Perner, 1983), and contrasted verbal vignettes about false beliefs (e.g., a protagonist has a false belief about an object's location; the critical condition) and linguistically matched vignettes about false physical states (physical representations depicting outdated scenes, e.g., a photograph showing an object that has since been removed; the control condition) (Figure 2) in a long-event-related design with a counterbalanced order across runs (when multiple runs were administered). Stimuli were presented one at a time. Participants read these vignettes and answered a true / false comprehension question after each one. Forty-one participants completed two runs and two completed one run due to time limitations, each lasting 262 s and consisting of 5 vignettes per condition.

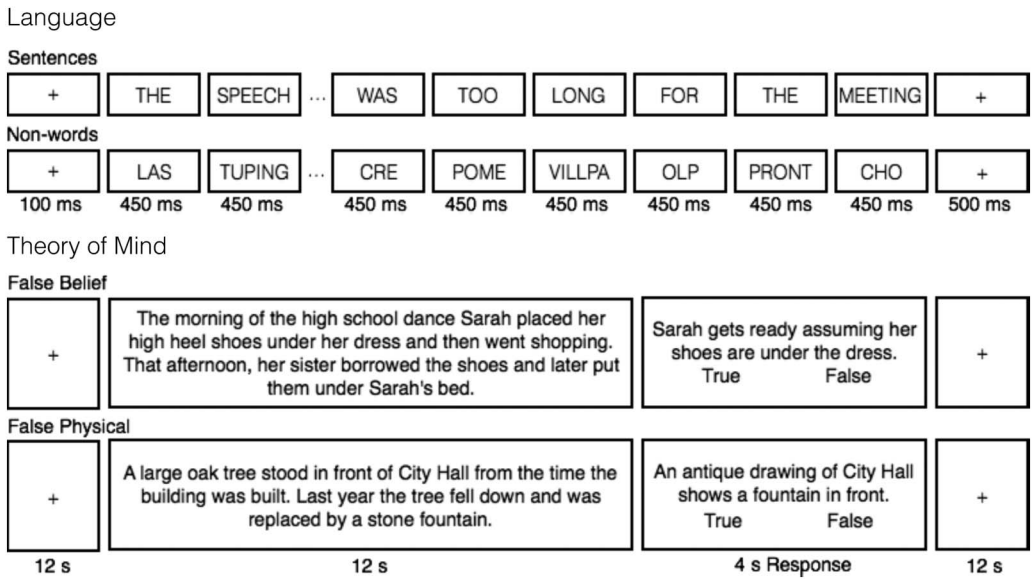

Figure 2. Sample trials from the functional localizer paradigms. *Language*: reading of sentences was contrasted with reading of sequences of pronounceable non-words (Fedorenko et al., 2010). *ToM*: reading of vignettes about false mental states was contrasted with reading of vignettes about false physical states, each followed by a true/false statement (Saxe & Kanwisher, 2003).

(A version of this localizer is available from https://saxelab.mit.edu/use-our-efficient-false-belief-localizer.)

An alternative, nonverbal ToM localizer task

One of the naturalistic conditions in this study (an animated short film, *Partly cloudy*; Reher & Sohn, 2009; see next section) has been previously used as a nonverbal ToM localizer (Jacoby et al., 2016; see also Richardson et al., 2018). To that end, it has been coded into *mental*, *physical*, *social*, and *pain* segments, and the regions defined by the mental > pain contrast have been validated against the traditional ToM localizer described above (see Jacoby et al., 2016, for details). Examples of mental content include a character falsely believing they have been abandoned by a companion, and a character observing others interacting happily after experiencing pain (4 events, 44 s total). Following a reviewer's suggestion, we used this localizer as an alternative ToM localizer in some of the analyses. We used the mental > physical contrast rather than mental > pain to maintain conceptual similarity with the verbal localizer's false belief > false physical contrast. The activations obtained with the two different control conditions were qualitatively similar. Examples of physical content include a wide shot of clouds and birds flying (3 events, 22 s total). The main goal was to ensure that the language–ToM dissociation is not due to an overly narrow definition of theory of mind in terms of false beliefs, implicit in the use of this particular type of mental state attribution in the standard, verbal localizer. Notably, many previous studies have shown that the ToM network defined with the false belief localizer responds to a wide range of mental state content besides (false) beliefs, including intentions, sources of evidence about others' minds, emotional pain, and the "minds" of group agents (e.g., Bruneau et al., 2012; Jenkins et al., 2014; Koster-Hale et al., 2014; Young & Saxe, 2008). Nevertheless, analyses that use the nonverbal ToM localizer should confirm that the results of the present study generalize beyond a particular way of localizing the ToM network.

MD localizer task (used in a replication analysis, as described below)

The task used to localize the MD network is described in detail in Fedorenko et al. (2011) and targets brain regions that support goal-directed effortful behaviors (e.g., Duncan, 2010, 2013). Briefly, we used a spatial working-memory task contrasting a harder version with an easier version (Supplemental Figure A1; Supporting Information can be found at https://doi.org/10.1162/nol_a_00071) in a blocked design with a counterbalanced condition order across runs (when multiple runs were administered). On each trial, participants saw a 3 × 4 grid and kept track of eight (hard version) or four (easy version) randomly generated locations that were sequentially flashed two at a time or one at a time, respectively. Then participants indicated their memory for these locations in a two-alternative, forced-choice paradigm via a button press, and received feedback. Of the 32 participants included in the replication analysis (i.e., non-overlapping with those used in the original study in Blank & Fedorenko, 2017), 23 participants completed two runs of the localizer, and 9 completed one run, each lasting 448 s and consisting of 6 blocks per condition and 4 fixation blocks.

The critical naturalistic task

In the main experiment, each participant listened to (over scanner-safe Sensimetrics headphones) and/or watched a set of naturalistic stimuli (varying between 4 min 50 s and 7 min 6 s in duration). Four of the conditions used linguistic materials: (i) a story (*Elvis* from the Natural Stories corpus; Futrell et al., 2021); (ii) an audio play (a segment from an HBO

miniseries, "Bad News" (audio only) from *Angels in America*; Kushner & Nichols, 2003); (iii) a naturalistic dialogue—a casual unscripted conversation between two female friends (recorded by JG); and (iv) a non-narrative expository text (a text about trees adapted from "Tree"; Wikipedia, n.d.) (recorded by JG). The first three of the linguistic conditions were rich in mental state content; the fourth was meaningful naturalistic discourse with little/no mental state content (see below for additional discussion). The three remaining conditions were videos with no linguistic content: (i) an animated short film (*Partly Cloudy* from Pixar; Reher & Sohn, 2009); (ii) a clip from a live action film ("Falling Asleep in Church"; Mr. Bean Official, 2009); and (iii) a custom-created Heider and Simmel style animation (Heider & Simmel, 1944) consisting of simple geometric shapes moving in ways so as to suggest intentional interactions designed to tell a story (e.g., a shape gets locked up inside a space, another shape goes on a quest to get help to release it, etc.). All three non-linguistic conditions were rich in mental state content. (Five additional conditions—included for some participants in another study—are of no interest to the current study.) All the materials are available on OSF (except in cases where copyright issues prevent us from doing so): https://osf.io/prghx/. In the resting state scan, used for one of our reality-check analyses, as described below, participants were instructed to close their eyes and let their mind wander but to remain awake while resting in the scanner for 5 min (the scanner lights were dimmed and the projector was turned off).

It is important to note that although we classify these naturalistic conditions into "types" in a binary way (i.e., either involving ToM or not, and either involving language or not), this should not be taken to suggest that there cannot be gradation within each category. Indeed, given the richness of the stimuli, there almost certainly is, at least for the ToM dimension. However, we do not pursue this question further given the small number of stimuli within each category, and their complexity. Another challenge is coding the materials, especially for mental state content. In particular, linguistically mediated mental state attribution often proceeds not from explicit mentions of mental states but from action descriptions, and in nonverbal settings ToM attribution proceeds exclusively from action observation. As a result, it is difficult to specify, especially in complex naturalistic materials, when mental state attribution is prompted. Conversely, ToM vocabulary need not lead to stronger mental state attribution (e.g., "Alan thinks that this is a nice house" vs. "Alan says that this is a nice house"; presumably, a mental state is ascribed to Alan in both cases). Thus, we do not attempt to quantify degrees of mental state attribution beyond the overall presence or (near-)absence of mental state content.

Another important point to acknowledge is that there may be a certain degree of "contamination" across categories. Specifically, as discussed in the Introduction, the −ToM condition (expository text) plausibly involves some degree of pragmatic inference, and such texts have previously been shown to elicit responses in parts of the ToM network (e.g., Ferstl & von Cramon, 2002; Jacoby & Fedorenko, 2020). In the case of language, even though binary classification is relatively straightforward (linguistic input is either present or not), one might argue that the language network may nevertheless be recruited to some extent owing to the communicative nature or rich semantics of the non-linguistic stimuli. Although prior work suggests that nonverbal communication does not recruit the language network (e.g., Deen et al., 2015; Jouravlev et al., 2019; Pritchett et al., 2018), some studies have found the language network is activated during the processing of visual event semantics (e.g., Ivanova et al., 2021). Critically for present purposes, however, to the extent that there is contamination in either direction, it should work *against* finding a language–ToM dissociation; i.e., our results might underestimate the true degree of dissociation. The rich and graded nature of the stimuli might help account for some results inconsistent with a *complete* dissociation, and we return to this question in the Discussion.

Data Acquisition and Preprocessing

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 echo-planar imaging (EPI) sequence with a 90° flip angle and using GRAPPA (generalized auto-calibrating partially parallel acquisitions) with an acceleration factor of 2; the following parameters were used: 31 4-mm-thick near-axial slices acquired in an interleaved order (with 10% distance factor), with an in-plane resolution of 2.1 × 2.1 mm, FoV in the phase encoding (A >> P) direction 200 mm and matrix size 96 × 96 mm, TR = 2,000 ms and TE = 30 ms. The first 10 s of each run were excluded to allow for steady-state magnetization.

Spatial preprocessing

Data preprocessing was performed with SPM12 software (https://www.fil.ion.ucl.ac.uk/spm/; using default parameters, unless specified otherwise) and supporting custom scripts in MATLAB (2020a; https://www.mathworks.com/). Preprocessing of anatomical data included normalization into a common space (Montreal Neurological Institute (MNI) template) and segmentation into probabilistic maps of the gray matter (GM), white matter (WM), and cerebro-spinal fluid (CSF). A GM mask was generated from the GM probability map, and resampled to 2 mm isotropic voxels to mask the functional data. The WM and CSF maps were used as described in temporal preprocessing below. Preprocessing of functional data included motion correction (realignment to the mean image using second-degree b-spline interpolation), normalization (estimated for the mean image using trilinear interpolation), resampling into 2 mm isotropic voxels, smoothing with a 4 mm FWHM (full width at half maximum) Gaussian filter and high-pass filtering at 200 s.

Temporal preprocessing

Additional preprocessing of data from the story comprehension runs was performed using the *Conn* toolbox (Whitfield-Gabrieli & Nieto-Castañón, 2012; https://www.nitrc.org/projects/conn) with default parameters, unless specified otherwise. BOLD signal time courses were extracted from WM and CSF. Five temporal principal components were obtained from each, as well voxel-wise averages. These were regressed out of each voxel's time course, along with additional noise regressors, specifically, six motion parameters estimated during off-line motion correction (three translations and three rotations) and their first temporal derivatives, and artifact time points (based on global signal outliers and motion). The residual signal was band-pass filtered (0.008–0.09 Hz) to preserve only low-frequency signal fluctuations (Cordes et al., 2001). This filtering did not influence the results reported below.

Participant-Specific Functional Localization of the Language and ToM (and MD, for a Replication Analysis) Networks

Modeling localizer data

For each localizer task, a standard mass univariate analysis was performed in SPM12 whereby a general linear model estimated the effect size of each condition in each experimental run.

These effects were each modeled with a boxcar function (representing entire blocks) convolved with the canonical hemodynamic response function. The model also included first-order temporal derivatives of these effects, as well as nuisance regressors representing entire experimental runs, off-line-estimated motion parameters, and time points classified as outliers (i.e., where the scan-to-scan differences in the global BOLD signal are above 5 standard deviations, or where the scan-to-scan motion is above 0.9 mm). The obtained weights were then used to compute the functional contrast of interest: for the language localizer, sentences > nonwords, for the ToM localizer false belief > false photo, and for the MD localizer (replication analysis; see the Stimuli and Procedure section), hard > easy spatial working memory.

Defining fROIs

Language and ToM (and MD, in the replication analysis) functional regions of interest (fROIs) were defined individually for each participant based on functional contrast maps from the localizer experiments (a toolbox for this procedure is available online; https://evlab.mit.edu/funcloc/). These maps were first restricted to include only GM voxels by excluding voxels that were more likely to belong to either the WM or the CSF based on SPM's probabilistic segmentation of the participant's structural data.

Then, fROIs in the language network were defined using group-constrained, participant-specific localization (Fedorenko et al., 2010). For each participant, the map of the sentences > nonwords contrast was intersected with binary masks that constrained the participant-specific language network to fall within areas where activations for this contrast are relatively likely across the population. These masks are based on a group-level representation of the contrast obtained from a previous sample of 220 participants. We used five such masks in the left hemisphere, including regions in the mid-to-posterior and anterior temporal lobe, as well as in the middle frontal gyrus, the inferior frontal gyrus, and its orbital part (Figure 4). A version of these masks is available online (https://evlab.mit.edu/funcloc/). In each of the resulting 5 masks, a participant-specific language fROI was defined as the top 10% of voxels with the highest contrast values. This top *n*% approach ensures that fROIs can be defined in every participant and that their sizes are the same across participants, allowing for generalizable results (Nieto-Castañón & Fedorenko, 2012).

For the ToM fROIs, we used masks derived from a group-level representation for the false belief > false physical contrast in an independent group of 462 participants (Dufour et al., 2013). These masks included regions in the left and right temporoparietal junction (L/RTPJ), precuneus / posterior cingulate cortex (L/RPC), and dorsal medial prefrontal cortex (MPFC; Figure 4). A version of these masks is available online (https://saxelab.mit.edu/use-our-theory-mind-group-maps), but the masks were edited as follows: The right superior temporal sulcus (RSTS) mask was excluded, as it covers the entire STS, which is known to show complex functional organization, with reduced ToM selectivity (Deen et al., 2015). The middle- and ventral-MPFC masks were also excluded to reduce the number of statistical comparisons in per-fROI analyses, but the dorsal MPFC and PC masks were split into left- and right-hemispheres, for a total of 6 masks.

Additionally, for the replication analysis, fROIs in the MD network were defined based on the hard > easy contrast in the spatial working memory task. Here, instead of using binary masks based on group-level functional data, we used anatomical masks (Blank et al., 2014; Blank & Fedorenko, 2017; Fedorenko et al., 2013; Tzourio-Mazoyer et al., 2002). Nine masks were used in each hemisphere, including regions in the middle frontal gyrus and its orbital

part, the opercular part of the inferior frontal gyrus, the precentral gyrus, the superior and inferior parts of the parietal lobe, the insula, the supplementary motor area, and the cingulate cortex (Supplemental Figure A2). (We note that functional masks derived for the MD network based on 197 participants were largely overlapping with the anatomical masks; we chose to use the anatomical masks to maintain comparability between our functional data and data from previous studies that have used these masks.)

In line with prior studies (e.g., Blank & Fedorenko, 2017; Blank et al., 2014; Paunov et al., 2019), the resulting fROIs showed small pairwise overlaps within individuals across networks, and overlapping voxels were excluded in fROI definition. In the current sample, the language-MD and ToM-MD overlaps were negligible, with a median overlap of 0 and average percentage overlap of fewer than 3% of voxels, on average across participants, relative to the total size of all fROIs in either network. Similarly, the language–ToM overlaps were small relative to all fROIs in either network (6.3% of voxels, on average across participants, relative to the total number of voxels in language fROIs and 3.2% relative to all ToM fROIs). This overlap was localized entirely to one pair of fROIs: the left posterior temporal (LPostTemp) language fROI and the left temporoparietal junction (LTPJ) ToM fROI, and was more substantial relative to the total sizes of just these two fROIs: 38.6 voxels, on average across participants, i.e., 13.1% out of 295 total LPostTemp voxels, and 11.6% out of 332 total LTPJ voxels. We therefore repeated all key analyses without excluding these voxels in defining the fROIs. The results of these alternative analyses were qualitatively and statistically similar.

Reality Check and Replication Analyses

Prior to performing our critical analyses, we conducted two reality-check analyses and—in line with increasing emphasis in the field on robust and replicable science (e.g., Poldrack et al., 2017)—an analysis aimed at replicating and extending a previous ISC-based finding from our lab (Blank & Fedorenko, 2017).

ISCs in perceptual cortices

Anatomical ROIs were additionally defined in early visual and auditory cortex in all participants. For visual cortex, regions included inferior, middle, and superior occipital cortex bilaterally (6 ROIs in total; masks available from https://www.nitrc.org/projects/wfu_pickatlas/). For auditory cortex, regions included posteromedial and anterolateral sections of Heschl's gyrus bilaterally (4 ROIs in total; Morosan et al., 2001; these regions are based on postmortem histology and have been used in a number of previous fMRI papers). All parcels used are available on OSF (https://osf.io/prghx/). Signal extraction, ISC estimation, and inferential statistics were performed identically to the critical analyses (see the Critical Analyses section). ISCs from these regions were used in a reality check (see the Results of Reality Check and Replication Analyses section), to ensure that a double-dissociation obtains between visually and auditorily presented conditions in these perceptual regions.

Resting state ISCs

A subset of 10 participants (who completed 1–7 of the critical conditions) completed a resting state scan, which was included to ensure that data acquisition, preprocessing, and modeling procedures do not induce spurious ISCs. To this end, signal extraction, ISC estimation, and inferential statistics were performed identically to the critical analyses (see the Critical Analyses section).

Replication analysis: Closer tracking of linguistic input by language regions than by domain-general MD regions

Blank and Fedorenko (2017) reported stronger ISCs during the processing of naturalistic linguistic materials in the language regions, compared to domain-general MD regions. The MD network has been implicated in executive processes and goal-directed behavior (e.g., Duncan, 2010, 2013), including in the domain of language (e.g., see Fedorenko, 2014, for a review; cf. Diachek et al., 2020; Fedorenko & Shain, 2021). We sought to replicate Blank and Fedorenko's key result in a new set of participants and to extend it to different types of linguistic materials. Specifically, the original study used narratives, including the narrative used in the present study along with three others. We expected the results to generalize to non-narrative linguistic conditions. Ten participants in our data set ($n = 6$ in the narrative condition) who also participated in the original study were excluded from this analysis. Again, signal extraction, ISC estimation, and inferential statistics were identical to the critical analyses (see the Critical Analyses section).

Critical Analyses

Computing ISCs

For each participant and fROI, BOLD signal time courses recorded during each naturalistic condition were extracted from each voxel beginning 6 s following the onset of the stimulus (to exclude an initial rise in the hemodynamic response relative to fixation, which could increase ISCs) and averaged across voxels. For each fROI, participant, and condition we computed an ISC value, namely, Pearson's product-moment correlation coefficient between the *z*-scored time course and the corresponding *z*-scored and averaged time course across the remaining participants (Lerner et al., 2011). ISCs were Fisher-transformed before statistical testing to improve normality (Silver & Dunlap, 1987).

Statistical testing

In each fROI, ISCs were then tested for significance against an empirical null distribution based on 1,000 simulated signal time courses that were generated by phase-randomization of the original data (Theiler et al., 1992). Namely, we generated null distributions for individual participants, fit each distribution with a Gaussian, and analytically combined the resulting parameters across participants. The true ISCs, also averaged across participants, were then *z*-scored relative to these empirical parameters and converted to one-tailed *p* values.

ISCs were compared across networks and condition types using LME regressions, implemented in MATLAB 2020a. ISCs were modeled with maximal random effects structure appropriate for each analysis (Barr et al., 2013; Baayen et al., 2008), including random intercepts for participants, with random slopes for the effects of interest, and crossed random intercepts for fROI and condition. Hypothesis testing was performed with two-tailed tests over the respective model coefficients, with Satterthwaite approximation for the degrees of freedom.

Further analyses performed within networks across condition types or against the theoretical null distribution (i.e., testing the intercept term), as well as those per fROI within a network or per condition across networks, also always included maximal random effects on the remaining grouping variables. The *p* values in these analyses are reported following false discovery rate (FDR) correction for multiple comparisons (Benjamini & Yekutieli, 2001).

Lastly, in comparisons against baseline, per-fROI analyses against empirical null distributions are also reported, which aim to ensure that, at the finest grain (each individual fROI and

condition, across participants), differences from baseline are independent of assumptions regarding data normality. These tests were also FDR corrected for multiple comparisons for all fROIs within a network, and across all seven conditions of interest.

RESULTS

Results of Reality Check and Replication Analyses

Reality check #1: ISCs in perceptual cortices

We examined ISCs for the conditions of interest in early auditory and visual cortex, grouping the conditions by presentation modality. As expected, we observed stronger ISCs for the auditory conditions in the auditory cortex, and stronger ISCs for the visual conditions in the visual cortex (Figure 3). The LME regression (see the Overall Experimental Design and Statistical Analyses section) revealed a strong crossover interaction (beta = 0.648, SE = 0.046, $t(89.79) = 13.970$, $p = 10^{-25}$).

Notably, we also observed that the visual areas weakly but reliably tracked the auditory conditions (beta = 0.066, SE = 0.012, $t(38.10) = 5.398$, $p = 10^{-7}$), and the early auditory areas reliably tracked the visual conditions (beta = 0.175, SE = 0.020, $t(24.15) = 8.573$, $p = 10^{-10}$). We return to the interpretation of these effects in the Discussion. For the time being, we note that care must be taken in interpreting deviations of the ISCs during "active" (cf. resting state) conditions from baseline.

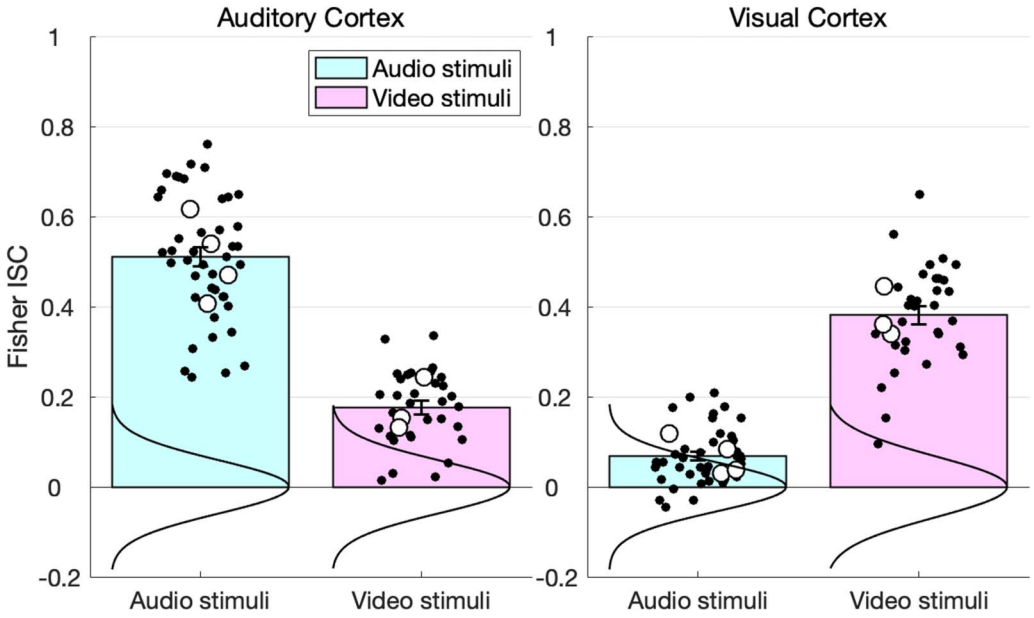

Figure 3. A reality-check analysis showing the expected double dissociation in inter-subject correlation (ISC) in perceptual (visual and auditory) cortices. Bars correspond to Fisher-transformed ISC coefficients (Pearson's r) in early visual and auditory cortex to the conditions of interest, grouped by modality of presentation (all linguistic [+Lang]: conditions: Story, audio play, dialogue, expository text were auditorily presented; the remaining conditions—animated short film, live action movie clip, and intentional shapes animation—were visually presented). Error bars are standard errors of the mean by participants. Black dots correspond to the individual participants' values. Large unfilled circles correspond to individual condition averages. Vertical curves are Gaussian fits to empirical null distributions. Stim, Stimuli.

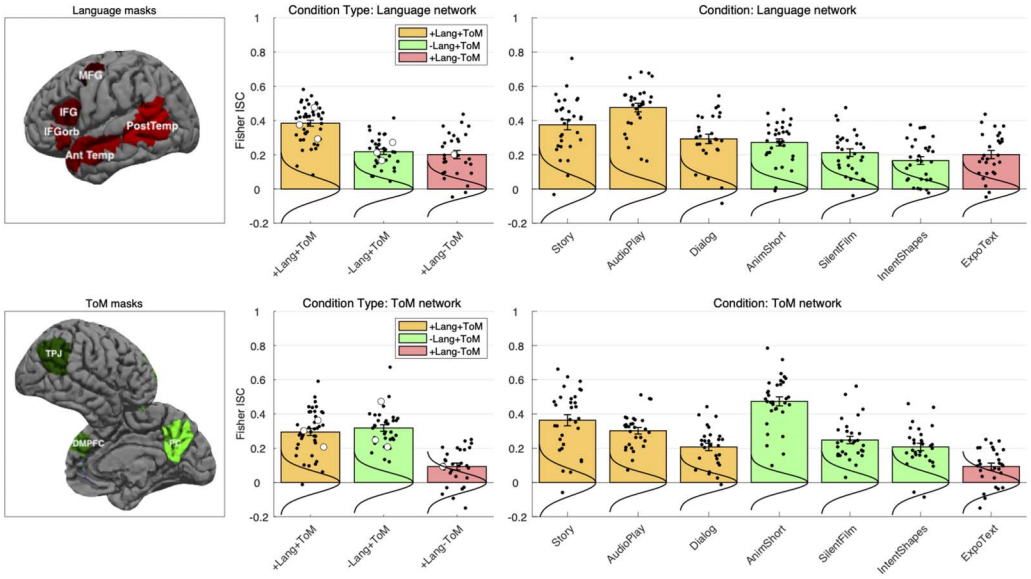

Figure 4. **Left**. Masks within which individual functional regions of interest (fROIs) were defined for each network: Language (Top, red): IFGorb, inferior frontal gyrus, orbital portion; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; AntTemp, anterior temporal cortex; PostTemp, posterior temporal cortex (only the classic left-hemisphere language regions were included in all analyses). ToM (Bottom, green): TPJ, temporoparietal junction; DMPFC, dorsomedial prefrontal cortex; PC, posterior cingulate cortex and precuneus. Both the right-hemisphere (shown) and left-hemisphere ToM regions were included (six regions total). **Middle**. Average inter-subject correlations (ISCs) per condition type in the language (top) and ToM (bottom) networks. Bars correspond to Fisher-transformed ISC coefficients (Pearson's *r*), averaged across regions of interest within each network, separately per condition. Colors represent condition types: +Lang +ToM, orange, −Lang +ToM, green, +Lang −ToM, red. Error bars are standard errors of the mean by participants. Black dots correspond to the individual participants' values. Large unfilled circles correspond to individual condition averages, shown individually in the right-most panels. Vertical curves are Gaussian fits to empirical null distributions. The key pattern is as follows: The ToM network tracks +ToM materials in both linguistic and non-linguistic conditions, but shows weak tracking of the non-ToM stimulus. The language network preferentially tracks linguistic materials over non-linguistic ones, and it tracks linguistic materials in both ToM and non-ToM conditions. **Right**. ISCs per individual naturalistic stimuli ("conditions"); conventions are as in middle panels.

Reality check #2: Resting state ISCs

To exclude the possibility that the ISCs in the critical analyses are driven by scanner noise or preprocessing/analysis procedures, we measured ISCs across a subset of 10 participants who were scanned in a 5-min resting state condition (Hasson et al., 2004). The ISCs during rest did not significantly differ from baseline in either the language or ToM networks, as assessed with an LME regression, or against the empirical null distribution. This analysis suggests that any above-baseline ISCs for our critical conditions are not an artifact of data acquisition, preprocessing, or analysis procedures.

Replication analysis: Closer tracking of linguistic input by language regions than by domain-general MD regions

We successfully replicated the key finding for the narrative condition (beta = 0.181, SE = 0.039, $t(34.89) = 4.664$, $p = 10^{-5}$; p values are FDR-corrected for the four linguistic conditions) and extended it to the audio play (beta = 0.307, $SE = 0.052$, $t(34.23) = 5.944$, $p = 10^{-7}$), the dialogue (beta = 0.157, SE = 0.039, $t(33.05) = 4.014$, $p = 10^{-5}$), and the expository text (beta = 0.129, SE = 0.036, $t(36.75) = 3.543$, $p = 10^{-4}$) conditions (Supplemental Figure A3). These results suggest that across diverse kinds of linguistic stimuli, the language network's

activity is more tightly coupled with the inputs, compared to the domain-general MD regions' activity.

Results of Critical Analyses

Evidence for dissociation between the language and ToM networks in direct network comparisons

We tested three key predictions, which, if supported, provide evidence in favor of language–ToM dissociability in naturalistic settings. First, the ToM network should track conditions rich in ToM content irrespective of whether these conditions are linguistic or non-linguistic, whereas the language network should track linguistic conditions more strongly than non-linguistic ones. Indeed, we found a network (language, ToM) × condition type (*linguistic*: narrative, audio play, dialogue, expository text; *non-linguistic*: animated film, live action film clip, Heider & Simmel-style animation) interaction (beta = 0.191, $SE = 0.059$, $t(84.06) = 3.270$, $p = 0.002$). We also found main effects of condition type and network: The linguistic conditions were—on average across networks—tracked more strongly than the non-linguistic conditions (beta = 0.124, $SE = 0.0459$, $t(96.47) = 2.700$, $p = 0.008$), and the language network tracked the conditions more strongly, on average, than the ToM network (beta = 0.096, $SE = 0.039$, $t(86.87) = 2.487$, $p = 0.015$).

Second, for non-linguistic conditions, the ToM network should exhibit stronger tracking of conditions with mental state content than the language network. Indeed, the ToM network showed higher ISCs than the language network (beta = 0.095, $SE = 0.043$, $t(35.59) = 2.197$, $p = 0.035$).

And third, for the linguistic condition without mental state content (the expository text), the language network should exhibit stronger tracking than the ToM network. Indeed, the language network showed higher ISCs than the ToM network (beta = 0.111, $SE = 0.044$, $t(19.10) = 2.507$, $p = 0.021$).

The same qualitative pattern obtains when the nonverbal ToM localizer is used to define the ToM fROIs (Supplemental Figure C1).

A more detailed characterization of the two networks' ISC profiles

In this section, we examine more closely the detailed pattern of ISCs in the two networks of interest. The first aim of these analyses is to establish that the observed dissociation is not driven by particular conditions or regions within the networks, but rather that different aspects of the data provide convergent support for the dissociation. The second aim is to highlight aspects of the ISC pattern that are not consistent with a complete language–ToM dissociation, and thus to evaluate the strength of counterevidence in favor of the null hypothesis, that language and ToM are not dissociable in naturalistic cognition (see Supplemental Figure B1 for ISCs per fROI for each network and condition).

ToM network

First, the ToM network reliably tracked each of the six conditions with mental state content: story (beta = 0.363, $SE = 0.037$, $t(26.55) = 9.693$, $p = 10^{-10}$; p values are FDR-corrected for seven conditions—we are including all conditions in the correction, not only the +ToM conditions), audio play (beta = 0.295, $SE = 0.028$, $t(10.54) = 10.716$, $p = 10^{-7}$), dialogue (beta = 0.203, $SE = 0.026$, $t(16.40) = 7.797$, $p = 10^{-7}$), animated short (beta = 0.472, $SE = 0.053$, $t(9.35) = 8.830$, $p = 10^{-6}$), live action film (beta = 0.241, $SE = 0.035$, $t(11.18) = 6.869$, $p = 10^{-6}$), and Heider & Simmel style animation (beta = 0.205, $SE = 0.026$, $t(15.72) = 7.838$, $p = 10^{-7}$). Moreover, in tests against the

empirical null distributions these effects were significant in every ToM fROI with the exception of the dialogue condition in the RPC, (all other *p*s < 0.04, FDR-corrected for the six fROIs and seven conditions).

Second, the ToM network showed no preference for linguistic vs. non-linguistic conditions with mental state content (beta = 0.018, *SE* = 0.043, *t*(49.74) = 0.675, *ns*), consistent with these regions' role in representing mental states irrespective of how this information is conveyed (e.g., Jacoby et al., 2016).

And third, the ToM network tracked the linguistic conditions with mental state content (story, audio play, dialogue) more strongly than the one without mental state content (expository text) (beta = 0.199, *SE* = 0.037, *t*(30.69) = 5.344, $p = 10^{-7}$), suggesting that the network represents mental state information in linguistic signals, rather than the linguistic signal itself. However, the ToM network did exhibit weaker but significantly above-baseline tracking of the expository text (beta = 0.093, *SE* = 0.022, *t*(22.22) = 4.165, *p* = 0.001). In per-fROI tests against the empirical null distributions, this effect was only reliable in the LTPJ (*p* = 0.015; all other *p*s > 0.05; FDR-corrected for the six fROIs). We consider possible explanations in the Discussion.

Language network

First, the language network reliably tracked each of the three linguistic conditions with mental state content: story (beta = 0.374, *SE* = 0.041, *t*(13.27) = 9.171, $p = 10^{-7}$; *p* values are FDR-corrected for seven conditions), audio play (beta = 0.479, *SE* = 0.046, *t*(8.75) = 10.370, $p = 10^{-6}$), and dialogue (beta = 0.294, *SE* = 0.045, *t*(9.72) = 6.491, $p = 10^{-4}$). Moreover, these effects were significant in every language fROI, in tests against the empirical null distributions (*p*s < 0.01, FDR-corrected for the five fROIs and seven conditions).

Second, importantly, the language network also reliably tracked the linguistic condition with no mental state content (beta = 0.204, *SE* = 0.044, *t*(8.72) = 4.668, $p = 10^{-4}$), and this effect, too, was significant in every language fROI (*p*s < 0.03, FDR-corrected for the five fROIs). This result suggests that mental state content is not necessary to elicit reliable ISCs in the language network.

And third, the language network showed stronger tracking of linguistic relative to non-linguistic conditions with mental state content (beta = 0.177, *SE* = 0.043, *t*(43.54) = 4.155, $p = 10^{-5}$). This result suggests a special role for linguistic input in driving the network's responses.

However, the language network exhibited some patterns that might be taken to suggest that mental state content—or social information more generally—is, to some extent, important for linguistic processing. First, the language network tracked the linguistic conditions with mental state content more strongly than the linguistic condition with no mental state content, i.e., the expository text (beta = 0.188, *SE* = 0.061, *t*(25.67) = 3.050, *p* = 0.005). This result may be taken to suggest that mental state content contributes to the language network's input tracking over and above the linguistic content alone. This interpretation warrants caution, however. In particular, reflecting the general challenges of naturalistic stimuli (see Discussion), the linguistic condition with no mental state content is not matched to the linguistic conditions with mental state content on various potentially relevant features, from how engaging they are, which could influence the depth of linguistic encoding, to specifically linguistic properties (e.g., lexical and syntactic complexity), which could also affect the strength of ISCs (e.g., Shain et al., 2020; Wehbe et al., 2021). Furthermore, only a single linguistic condition with no mental state content was included in the current study, making it difficult to rule out idiosyncratic features driving the difference.

And second, the above-baseline ISCs in the language network for the non-linguistic conditions—although weaker than those for the linguistic conditions—are also notable, suggesting some degree of reliable tracking in the language network for non-linguistic meaningful information (see also Ivanova et al., 2021, for evidence of reliable responses in the language network to visual events).

DISCUSSION

Much prior work in cognitive neuroscience has suggested—based on traditional controlled experimental paradigms—that the network of brain regions that support linguistic interpretation and the brain regions that support mental state reasoning are distinct (e.g., Deen et al., 2015; Fedorenko et al., 2011; Mar, 2011; Mason & Just, 2009; Paunov, 2019; Saxe & Kanwisher, 2003). However, such paradigms differ drastically from real-world cognition, where we process rich and complex information. And linguistic and social cognition seem to be strongly intertwined in everyday life. Here, we tested whether the language and ToM networks are dissociated in their functional profiles as assessed using the ISC approach, where neural activity patterns are correlated across individuals during the processing of naturalistic materials (e.g., Hasson et al., 2004, 2008). Following Blank and Fedorenko (2017), we combined the ISC approach with the power of individual-participant functional localization (e.g., Brett et al., 2002; Fedorenko et al., 2010; Nieto-Castañón & Fedorenko, 2012; Saxe, Brett, & Kanwisher, 2006). This synergistic combination has two key advantages over the whole-brain voxel-wise ISC approach, where individual brains are first anatomically aligned and, then, each stereotaxic location serves as a basis for comparing signal time courses across participants. First, relating the resulting cortical topography of ISCs to the topography of known functional brain networks can only proceed through "reverse inference" based on anatomy (Fedorenko, 2021; Poldrack, 2006). Instead, evaluating signal time courses from functionally defined regions ensures interpretability and allows us to straightforwardly link our findings to the wealth of prior studies characterizing the response profiles of our two networks of interest. And second, this approach allows us to directly test the correlations in the language network against those in the ToM network. Such an explicit comparison between networks allows for stronger inferences compared with those licensed when each network is separately tested against a null baseline and differences across networks are indirectly inferred (e.g., see Nieuwenhuis et al., 2011, for discussion).

We examined the ISCs in the language and ToM networks during the processing of seven naturalistic conditions: three linguistic conditions with mental state content (+linguistic, +ToM), three non-linguistic conditions (silent animations and live action films) with social content but no language (–linguistic, +ToM), and a linguistic condition with no social content (+linguistic, −ToM). We found reliable differences in the ISC patterns between the language and ToM networks, in support of the hypothesis that language and ToM are dissociable even during the processing of rich and complex naturalistic materials. In particular, the ToM network tracked materials rich in mental state content irrespective of whether this content was presented linguistically or non-linguistically (see also Jacoby et al., 2016), but it showed only weak tracking of the stimulus with no mental state content. In contrast, the language network preferentially tracked linguistic materials over non-linguistic ones, and it did so regardless of whether these materials contained information about mental states.

These results expand on the existing body of knowledge about language and social cognition, with both theoretical and methodological implications. Critically, the observed dissociation extends prior findings of dissociable functional profiles between the language and the

ToM networks during task-based paradigms to rich naturalistic conditions. This result suggests that the two networks represent *different kinds of information*. (They may also perform *distinct computations* on the perceptual inputs, though the idea of a canonical computation carried out across the cortex is gaining ground (e.g., Fedorenko & Shain, 2021; Keller & Mrsic-Flogel, 2018), and predictive processing seems like one likely candidate (e.g., Koster-Hale & Saxe, 2013; Shain et al., 2020). In particular, the language regions appear to track linguistic features of the input (see also Shain et al., 2020, 2021; Wehbe et al., 2021). Our results extend prior findings from ISC paradigms (Blank & Fedorenko, 2017; Honey et al., 2012; Lerner et al., 2011; Regev et al., 2013; Schmälzle et al., 2015; Silbert et al., 2014; Wilson et al., 2007), which all used materials rich in mental state content, as is typical of linguistic information, to a stimulus that is largely devoid of information about mental states—an expository text. Strong tracking of the latter stimulus aligns with prior findings from task-based paradigms of robust responses to linguistic materials with little or no mental state content (e.g., Deen et al., 2015; Jacoby & Fedorenko, 2020).

The ToM regions, in contrast, appear to track some features related to representing mental states across diverse kinds of representations (linguistic materials, animations, including highly abstract and minimalistic ones, and live action movies), again aligning with prior findings from task-based paradigms (e.g., Castelli et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000; Jacoby et al., 2016; Ruby & Decety, 2003; Saxe & Kanwisher, 2003; Saxe, Schulz, & Jiang, 2006; Vogeley et al., 2001). It is important to keep in mind that the fact that the two networks are dissociable does not imply that they do not interact. Indeed, Paunov et al. (2019) reported reliably above-chance correlations in the patterns of inter-regional synchronization between the language and ToM networks, suggesting some degree of functional integration.

On the methodological level, these results vindicate the divide-and-conquer strategy in general, where cognitive domains are treated as components of a "nearly decomposable system" (Simon, 1962)—and the functional localization approach (e.g., Brett et al., 2002; Fedorenko et al., 2010; Saxe, Brett, & Kanwisher, 2006) in particular. Language and ToM appear to be distinct, supported by dissociable cortical networks for the processing of linguistic vs. mental state information, at least in adulthood (see also Braga et al., 2020). It is therefore justifiable to study each cognitive faculty and each network separately, although further probing the mechanisms of their potential interactions is equally important.

Although the overall pattern clearly supports a language–ToM dissociation between the two networks, some aspects of the results are not in line with a complete dissociation. In particular, (1) the language regions show reliable tracking of non-linguistic conditions with mental state content; (2) the language regions show stronger tracking of linguistic conditions with than without mental state content; and (3) the ToM regions show weak but reliable tracking of the linguistic condition with no mental state content. These findings may be due to methodological limitations: The above-baseline ISCs (especially the relatively weak ones) may not reflect stimulus tracking. Although we have ruled out the possibility that the above-baseline ISCs are driven by acquisition, preprocessing, or analysis artifacts in our reality-check analysis of resting state data, they could be driven by other, non-mutually-exclusive, factors. One possibility is that inter-network interactions could induce ISCs. In particular, given that the language and the ToM network show some degree of synchronization in activity during naturalistic cognition (Paunov et al., 2019), the ToM network's tracking of a linguistic stimulus with no mental state content, for example, may be due to the fact that the language system is tracking this stimulus, and there is some "leakage" of this tracking to the ToM network through inter-network synchronization. Similarly, the language network's preference of linguistic

stimuli with mental state content over those without mental state content may be due to the leakage from the ToM network.

Another possibility is that the incomplete dissociation between language and ToM (at least with respect to (1) and (3) above) may be at least partly attributable to pragmatic processing, arguably present across all conditions: The non-linguistic ToM conditions are still story-like and hence communicative, and the linguistic non-ToM stimulus arguably still requires attribution of communicative intentions, as discussed in the introduction. The division of labor between the language and ToM networks in pragmatic inference is an exciting future direction. Demonstrating that these two networks are, in the first place, dissociable, even during rich naturalistic cognition—the goal of the present study—is an important step to pursuing this line of research. With this groundwork in place, neuroimaging evidence can be increasingly brought to bear on the question of which aspects of pragmatics require mental state reasoning, as evidenced by the engagement of the ToM network. Given the broad scope of pragmatics, encompassing diverse heterogeneous phenomena, an empirically motivated pragmatic taxonomy may be developed by investigating whether some classes of pragmatic inference are resolved within the language network proper (e.g., "lower-level" inferences about lexical or syntactic ambiguity) whereas others (e.g., establishing discourse coherence or understanding irony) require the ToM network (e.g., Bosco et al., 2018; Hagoort & Levinson, 2014; Sperber & Wilson, 2002).

Finally, the stimuli themselves may be too confounded to fully dissociate the respective contributions of language- and ToM-related components. For example, in linguistic stimuli, mental state attribution often requires particular syntactic structures—sentential complements.

More generally, the use of naturalistic materials, despite their advantages (e.g., Hasson et al., 2018; Nastase et al., 2020; Sonkusare et al., 2019), is associated with a host of challenges. The key one, mentioned above, is that certain features are necessarily confounded in naturalistic settings and can only be dissociated through careful experimentation and altering the natural statistics of the input. Relying on naturalistic materials alone can lead to wrong conclusions about the cognitive and neural architecture. This problem is especially pronounced in studying the relationship between language and social cognition given that language is primarily used in social settings and to share socially relevant information. The use of linguistic materials with no social information and of non-linguistic mental-state-rich materials has been critical, here and in earlier studies, to uncover the dissociation that holds between the language and ToM systems.

Another challenge associated with the use of naturalistic materials is that they are difficult, or altogether impossible, to match for diverse properties bound to affect neural responses. Again, this problem presents a particular challenge in comparing responses to materials rich in social information vs. materials devoid of such information given that the former are, almost by definition, going to be more engaging and exciting given the social nature of primates, including humans (e.g., Aronson, 1980; Cheney & Seyfarth, 1990; Tomasello, 2014). Possible ways to address these concerns could involve (a) characterizing the natural statistics of co-occurrences between linguistic and social processing in order to better understand how well the naturalistic stimuli reflect those statistics and perhaps altering naturalistic conditions to allow dissociating features that commonly co-occur in life; (b) developing novel neural analysis methods to isolate the components of neural signals attributable to a particular cognitive process / brain network (e.g., using analytic methods well suited to high-dimensional data such as independent components analysis (ICA) across both cortical space and large feature spaces representative of naturalistic environments (e.g., Norman-Haignere et al., 2015), or across both cortical space and time (e.g., probabilistic ICA; see Beckmann et al., 2005, for

an overview); and (c) carefully annotating naturalistic materials and performing reverse correlation analyses (e.g., Hasson et al., 2004; Richardson et al., 2018) in an effort to understand the precise features that elicit increases in neural responses in different brain regions. The latter approach may be particularly informative with respect to the question of possible *gradation* of demands on language and ToM processing both within naturalistic stimuli and across conditions that we here grouped in the same type or general categories of +/−ToM and +/−Lang. Our results seem to suggest that considerable within-category heterogeneity exists in the degree of stimulus tracking (e.g., in the language network, the dialogue is tracked less strongly than the narrative, and in the ToM network, the Heider & Simmel-style animation is tracked less strongly than the animated film). Our data set is not ideally suited for such investigation because it includes only a single instance of each condition, and no attempt was made to match the conditions on at least *some* dimensions that may improve homogeneity, but this is a promising direction for future work.

A few other exciting future directions are worth highlighting, some of which may build directly on the approaches introduced and the findings reported in the current study. First, the language and ToM networks appear to be dissociable in the adult mind and brain. However, it is possible—perhaps even plausible—that this dissociation emerges over the course of development. Prior neuroimaging work has shown that the ToM network becomes gradually more specialized for mental state attribution (e.g., Gweon et al., 2012; Richardson et al., 2018; Saxe et al., 2009); and this specialization appears to be protracted with delayed language acquisition (Richardson et al., 2020). Very little is known about how specialization for linguistic processing (e.g., Fedorenko et al., 2011; Monti et al., 2012) emerges. Perhaps early on in development, a set of lateral frontal, lateral temporal, and midline cortical areas are tuned to *any* socially relevant information, and these areas later fractionate into those specialized for processing linguistic signals vs. those for mental state attribution vs. those that support many other kinds of social signals, both visual (e.g., eye gaze, facial expressions, gestures) and auditory (e.g., nonverbal vocalizations, prosodic information, speech acoustics). This fractionation is likely driven by computational and metabolic advantages of localized processing (e.g., Barlow, 1995; Chklovskii & Koulakov, 2004; Foldiak & Young, 1995; Olshausen & Field, 2004; see Kanwisher, 2010, for discussion). Probing linguistic and social cognition across the lifespan will be critical to understand how the two networks form and develop, leading to the segregation we observe in the adult brain.

Second, as noted above, given the likely frequent interactions between the language and ToM networks (including their most strongly dissociated components), searching for possible mechanisms of those interactions (e.g., Paunov et al., 2019) seems critical. This would require a combination of studies characterizing the patterns of anatomical connections for the language and ToM regions (e.g., Saur et al., 2010; Wiesmann et al., 2017) and studies probing online interactions using methods with high temporal resolution, like magnetoencephalography or intracranial recordings.

Third, we are still a long way away from a mechanistic-level understanding of what the language or the ToM regions do. The use of naturalistic stimuli, including in the context of the ISC approach, is promising. In particular, by examining the points in the stimulus where most participants show increases in neural activity can help generate (and subsequently test) specific hypotheses about the necessary and sufficient features of the input that are required to elicit neural responses in the relevant brain regions.

To conclude, we have demonstrated that the dissociation between the language and ToM networks that has been previously reported based on traditional task paradigms robustly

generalizes to rich naturalistic conditions. However, the precise nature of each network's representations and computations, the emergence of these networks in development, and the mechanisms for information sharing between them remain to be discovered.

FUNDING INFORMATION

Evelina Fedorenko, National Institute on Deafness and Other Communication Disorders (https://dx.doi.org/10.13039/100000055), Award ID: DC016607. Evelina Fedorenko, National Institute on Deafness and Other Communication Disorders (https://dx.doi.org/10.13039/100000055), Award ID: DC016950. Evelina Fedorenko, Simons Foundation (https://dx.doi.org/10.13039/100000893).

AUTHOR CONTRIBUTIONS

Alexander M. Paunov: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft. **Idan A. Blank**: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization. **Olessia Jouravlev**: Data curation; Investigation. **Zachary Mineroff**: Data curation; Formal analysis; Investigation; Software. **Jeanne Gallée**: Data curation; Resources. **Evelina Fedorenko**: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – review & editing.

REFERENCES

Amalric, M., & Dehaene, S. (2018). Cortical circuits for mathematical knowledge: Evidence for a major subdivision within the brain's semantic networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), Article 20160515. https://doi.org/10.1098/rstb.2016.0515, PubMed: 29292362

Anderson, A. J., Lalor, E. C., Lin, F., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Grimm, S., & Wang, X. (2019). Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cerebral Cortex*, *29*(6), 2396–2411. https://doi.org/10.1093/cercor/bhy110, PubMed: 29771323

Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. (2006). Intact first- and second-order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience*, *1*(3–4), 334–348. https://doi.org/10.1080/17470910601038693, PubMed: 18633798

Aronson, E. (1980). *The social animal*. Palgrave Macmillan.

Ayyash, D., Malik-Moraleda, S., Gallée, J., Affourtit, J., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). The universal language network: A cross-linguistic investigation spanning 45 languages and 11 language families. *BioRxiv*. https://doi.org/10.1101/2021.07.28.454040

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bahlmann, J., Rodriguez-Fornells, A., Rotte, M., & Münte, T. F. (2007). An fMRI study of canonical and noncanonical word order in German. *Human Brain Mapping*, *28*(10), 940–949. https://doi.org/10.1002/hbm.20318, PubMed: 17274018

Barlow, H. (1995). The neuron doctrine in perception. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 415–435). MIT Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001, PubMed: 24403724

Bautista, A., & Wilson, S. M. (2016). Neural responses to grammatically and lexically degraded speech. *Language, Cognition and Neuroscience*, *31*(4), 567–574. https://doi.org/10.1080/23273798.2015.1123281, PubMed: 27525290

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1457), 1001–1013. https://doi.org/10.1098/rstb.2005.1634, PubMed: 16087444

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bischetti, L., Ceccato, I., Lecce, S., Cavallini, E., & Bambini, V. (2019). Pragmatics and theory of mind in older adults' humor comprehension. *Current Psychology*, *2019*, 1–17. https://doi.org/10.1007/s12144-019-00295-w

Blank, I. [A.], Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069, PubMed: 26666896

Blank, I. A., & Fedorenko, E. (2017). Domain-general brain regions do not track linguistic input as closely as language-selective regions. *Journal of Neuroscience*, *37*(41), 9999–10011. https://doi.org/10.1523/JNEUROSCI.3642-16.2017, PubMed: 28871034

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, Article 116925. https://doi.org/10.1016/j.neuroimage.2020.116925, PubMed: 32407994

Blank, I. [A.], Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013, PubMed: 24872535

Bosco, F. M., Tirassa, M., & Gabbatore, I. (2018). Why pragmatics and theory of mind do not (completely) overlap. *Frontiers in Psychology*, *9*, 1453. https://doi.org/10.3389/fpsyg.2018.01453, PubMed: 30150960

Braga, R. M., DiNicola, L. M., Becker, H. C., & Buckner, R. L. (2020). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *Journal of Neurophysiology*, *124*(5), 1415–1448. https://doi.org/10.1152/jn.00753.2019, PubMed: 32965153

Brett, M., Johnsrude, I. S., & Owen, A. M. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, *3*(3), 243–249. https://doi.org/10.1038/nrn756, PubMed: 11994756

Bruneau, E. G., Pluta, A., & Saxe, R. (2012). Distinct roles of the "shared pain" and "theory of mind" networks in processing others' emotional suffering. *Neuropsychologia*, *50*(2), 219–231. https://doi.org/10.1016/j.neuropsychologia.2011.11.008, PubMed: 22154962

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*(3), 314–325. https://doi.org/10.1006/nimg.2000.0612, PubMed: 10944414

Chen, G., Taylor, P. A., & Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *NeuroImage*, *147*, 952–959. https://doi.org/10.1016/j.neuroimage.2016.09.066, PubMed: 27729277

Cheney, D., & Seyfarth, R. (1990). Attending to behaviour versus attending to knowledge: Examining monkeys' attribution of mental states. *Animal Behaviour*, *40*(4), 742–753. https://doi.org/10.1016/S0003-3472(05)80703-1

Chklovskii, D. B., & Koulakov, A. A. (2004). Maps in the brain: What can we learn from them? *Annual Review of Neuroscience*, *27*, 369–392. https://doi.org/10.1146/annurev.neuro.27.070203.144226, PubMed: 15217337

Cordes, D., Haughton, V. M., Arfanakis, K., Carew, J. D., Turski, P. A., Moritz, C. H., Quigley, M. A., & Meyerand, M. E. (2001). Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data. *American Journal of Neuroradiology*, *22*(7), 1326–1333. PubMed: 11498421

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111, PubMed: 26048954

Diachek, E., Blank, I., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. *Journal of Neuroscience*, *40*(23), 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020, PubMed: 32317387

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D. E., & Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLOS ONE*, *8*(9), Article e75468. https://doi.org/10.1371/journal.pone.0075468, PubMed: 24073267

Dunbar, R. I. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, *8*(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

Dunbar, R. I., Marriott, A., & Duncan, N. D. (1997). Human conversational behavior. *Human Nature*, *8*(3), 231–246. https://doi.org/10.1007/BF02912493, PubMed: 26196965

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004, PubMed: 20171926

Duncan, J. (2013). The structure of cognition: Attentional episodes in mind and brain. *Neuron*, *80*(1), 35–50. https://doi.org/10.1016/j.neuron.2013.09.015, PubMed: 24094101

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, *5*, Article 335. https://doi.org/10.3389/fpsyg.2014.00335, PubMed: 24803909

Fedorenko, E. (2021). The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences*, *40*, 105–112. https://doi.org/10.1016/j.cobeha.2021.02.023

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433. https://doi.org/10.1073/pnas.1112937108, PubMed: 21885736

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, Article 104348. https://doi.org/10.1016/j.cognition.2020.104348, PubMed: 32569894

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110, PubMed: 24062451

Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010, PubMed: 20410363

Fedorenko, E., Nieto-Castañón, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014, PubMed: 21945850

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113, PubMed: 27671642

Fedorenko, E., & Shain, C. (2021). Similarity of computations across domains does not imply shared implementation: The case of language comprehension. *Current Directions in Psychological Science*, *30*(6), 526–534. https://doi.org/10.1177/09637214211046955, PubMed: 35295820

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, *1369*(1), 132–153. https://doi.org/10.1111/nyas.13046, PubMed: 27096882

Feng, W., Wu, Y., Jan, C., Yu, H., Jiang, X., & Zhou, X. (2017). Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain and Language*, *171*, 52–61. https://doi.org/10.1016/j.bandl.2017.04.005, PubMed: 28527316

Feng, W., Yu, H., & Zhou, X. (2021). Understanding particularized and generalized conversational implicatures: Is theory-of-mind necessary? *Brain and Language*, *212*, Article 104878. https://doi.org/10.1016/j.bandl.2020.104878, PubMed: 33096372

Ferstl, E. C., & von Cramon, D. Y. (2002). What does the frontomedian cortex contribute to language processing: Coherence or theory of mind? *NeuroImage*, *17*(3), 1599–1612. https://doi.org/10.1006/nimg.2002.1247, PubMed: 12414298

Finlayson, M. A., & Winston, P. H. (2011). Narrative is a key cognitive competency. In *BICA 2011: Proceedings of the 2nd annual international conference on biologically inspired cognitive architectures* (p. 110). IOS Press.

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T., Mohlberg, H., Amunts, K., & Zilles, K. (2008). Cortical folding patterns and predicting cytoarchitecture. *Cerebral Cortex*, *18*(8), 1973–1980. https://doi.org/10.1093/cercor/bhm225, PubMed: 18079129

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R, PubMed: 8556839

Fodor, J. A. (1983). *The modularity of mind*. MIT Press. https://doi.org/10.7551/mitpress/4737.001.0001

Foldiak, P., & Young, M. (1995). Sparse coding in the primate cortex. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 895–898). MIT Press.

Frith, U., & Happé, F. (1994). Autism: Beyond "theory of mind." *Cognition*, *50*(1–3), 115–132. https://doi.org/10.1016/0010-0277(94)90024-8, PubMed: 8039356

Frost, M. A., & Goebel, R. (2012). Measuring structural–functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. *NeuroImage*, *59*(2), 1369–1381. https://doi.org/10.1016/j.neuroimage.2011.08.035, PubMed: 21875671

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The natural stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources & Evaluation*, *55*, 63–77. https://doi.org/10.1007/s10579-020-09503-7, PubMed: 34720781

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of "theory of mind" in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11–21. https://doi.org/10.1016/S0028-3932(99)00053-6, PubMed: 10617288

Gallivan, J. P., Cavina-Pratesi, C., & Culham, J. C. (2009). Is that within reach? fMRI reveals that the human superior parieto-occipital cortex encodes objects reachable by the hand. *Journal of Neuroscience*, *29*(14), 4381–4391. https://doi.org/10.1523/JNEUROSCI.0377-09.2009, PubMed: 19357266

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. In J. Kulas, J. H. Fetzer, & T. L. Rankin (Eds.), *Philosophy, language, and artificial intelligence* (pp. 49–66). Springer. https://doi.org/10.1007/978-94-009-2727-8_2

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003

Grice, H. P. (1991). Meaning. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 213–223). Harvard University Press. (Meaning originally published in 1957)

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development*, *83*(6), 1853–1868. https://doi.org/10.1111/j.1467-8624.2012.01829.x, PubMed: 22849953

Hagoort, P., & Levinson, S. C. (2014). Neuropragmatics. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The cognitive neurosciences* (pp. 667–674). MIT Press.

Hammer, A., Jansma, B., Tempelmann, C., & Münte, T. F. (2011). Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology*, *2*, Article 32. https://doi.org/10.3389/fpsyg.2011.00032, PubMed: 21713189

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*, 135–157. https://doi.org/10.1016/j.cognition.2018.06.018, PubMed: 30053570

Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron*, *57*(3), 452–462. https://doi.org/10.1016/j.neuron.2007.12.009, PubMed: 18255037

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48. https://doi.org/10.1016/j.tics.2009.10.011, PubMed: 20004608

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506, PubMed: 15016991

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not lost in translation: Neural responses shared across languages. *Journal of Neuroscience*, *32*(44), 15277–15283. https://doi.org/10.1523/JNEUROSCI.1800-12.2012, PubMed: 23115166

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637, PubMed: 27121839

Ivanova, A. A., Mineroff, Z., Zimmerer, V., Kanwisher, N., Varley, R., & Fedorenko, E. (2021). The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, *2*(2), 176–201. https://doi.org/10.1162/nol_a_00030

Ivanova, A., Siegelman, M., Cheung, C., Pongos, A., Kean H., & Fedorenko, E. (2020, October 21–24). *Effect of task on sentence processing* [Poster presentation]. Twelfth Annual Meeting of the Society for the Neurobiology of Language Conference (held online).

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U.-M., Bers, M. U., & Fedorenko, E. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *Elife*, *9*, Article e58906. https://doi.org/10.7554/eLife.58906, PubMed: 33319744

Jackendoff, R. S. (2009). *Language, consciousness, culture: Essays on mental structure*. MIT Press.

Jacoby, N., Bruneau, E., Koster-Hale, J., & Saxe, R. (2016). Localizing pain matrix and theory of mind networks with both verbal and non-verbal stimuli. *NeuroImage*, *126*, 39–48. https://doi.org/10.1016/j.neuroimage.2015.11.025, PubMed: 26589334

Jacoby, N., & Fedorenko, E. (2020). Discourse-level comprehension engages medial frontal theory of mind brain regions even for expository texts. *Language, Cognition and Neuroscience*, *35*(6), 780–796. https://doi.org/10.1080/23273798.2018.1525494, PubMed: 32984430

Jang, G., Yoon, S. A., Lee, S. E., Park, H., Kim, J., Ko, J. H., & Park, H. J. (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage*, *81*, 61–72. https://doi.org/10.1016/j.neuroimage.2013.05.027, PubMed: 23684863

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, *21*(12), 2434–2444. https://doi.org/10.1162/jocn.2008.21179, PubMed: 19199402

Jenkins, A. C., Dodell-Feder, D., Saxe, R., & Knobe, J. (2014). The neural bases of directed and spontaneous mental state attributions to group agents. *PLOS ONE*, *9*(8), Article e105341. https://doi.org/10.1371/journal.pone.0105341, PubMed: 25140705

Jouravlev, O., Zheng, D., Balewski, Z., Pongos, A. L. A, Levan, Z., Goldin-Meadow, S., & Fedorenko, E. (2019). Speech-accompanying gestures are not processed by the language-processing mechanisms. *Neuropsychologia*, *132*, Article 107132. https://doi.org/10.1016/j.neuropsychologia.2019.107132, PubMed: 31276684

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, *60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055, PubMed: 22398396

Kandylaki, K. D., Nagels, A., Tune, S., Wiese, R., Bornkessel-Schlesewsky, I., & Kircher, T. (2015). Processing of false belief passages during natural story comprehension: An fMRI study. *Human Brain Mapping*, *36*(11), 4231–4246. https://doi.org/10.1002/hbm.22907, PubMed: 26356583

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107, PubMed: 20484679

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003, PubMed: 30359606

Klepousniotou, E., Gracco, V. L., & Pike, G. B. (2014). Pathways to lexical ambiguity: fMRI evidence for bilateral fronto-parietal involvement in language processing. *Brain and Language*, *131*, 56–64. https://doi.org/10.1016/j.bandl.2013.06.002, PubMed: 24183467

Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition*, *133*(1), 65–78. https://doi.org/10.1016/j.cognition.2014.04.006, PubMed: 24960530

Koster-Hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & M. V. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 132–163). https://doi.org/10.1093/acprof:oso/9780199692972.003.0009

Kushner, T. (Writer), & Nichols, M. (Director). (2003, December 7). Bad news (Episode 1, Season 1; Audio only) [HBO miniseries episode]. In C. D. Costas (Producer), *Angels in America*. Avenue Pictures, HBO Films. Retrieved from https://www.imdb.com/title/tt0318997/

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is theory of mind associated with pragmatics in middle childhood? *Journal of Child Language*, *46*(2), 393–407. https://doi.org/10.1017/S030500091800048X, PubMed: 30442207

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011, PubMed: 21414912

Liu, Y. F., Kim, J., Wilson, C., & Bedny, M. (2020). Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *Elife*, *9*, Article e59340. https://doi.org/10.7554/eLife.59340, PubMed: 33319745

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, *62*, 103–134. https://doi.org/10.1146/annurev-psych-120709-145406, PubMed: 21126178

Mason, R. A., & Just, M. A. (2009). The role of the theory-of-mind cortical network in the comprehension of narratives. *Language and Linguistics Compass*, *3*(1), 157–174. https://doi.org/10.1111/j.1749-818X.2008.00122.x, PubMed: 19809575

Mason, R. A., Just, M. A., Keller, T. A., & Carpenter, P. A. (2003). Ambiguity in the brain: What brain imaging reveals about the processing of syntactically ambiguous sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1319–1338. https://doi.org/10.1037/0278-7393.29.6.1319, PubMed: 14622064

McMillan, C. T., Clark, R., Gunawardena, D., Ryant, N., & Grossman, M. (2012). fMRI evidence for strategic decision-making during resolution of pronoun reference. *Neuropsychologia*, *50*(5), 674–687. https://doi.org/10.1016/j.neuropsychologia.2012.01.004, PubMed: 22245014

Mellem, M. S., Jasmin, K. M., Peng, C., & Martin, A. (2016). Sentence processing in anterior superior temporal cortex shows a social-emotional bias. *Neuropsychologia*, *89*, 217–224. https://doi.org/10.1016/j.neuropsychologia.2016.06.019, PubMed: 27329686

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought beyond language neural dissociation of algebra and natural language. *Psychological Science*, *23*(8), 914–922. https://doi.org/10.1177/0956797612437427, PubMed: 22760883

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage*, *13*(4), 684–701. https://doi.org/10.1006/nimg.2000.0715, PubMed: 11305897

Mr. Bean Official. (2009, August 25). Falling asleep in church (Funny Clip) [Video file]. Retrieved from https://www.youtube.com/watch?v=bh__g-ZZ6WA

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, Article 117254. https://doi.org/10.1016/j.neuroimage.2020.117254, PubMed: 32800992

Nieto-Castañón, A., & Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*, *63*(3), 1646–1669. https://doi.org/10.1016/j.neuroimage.2012.06.065, PubMed: 22784644

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886, PubMed: 21878926

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *8*(6), 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035, PubMed: 26687225

Nowak, M. [A.], & Highfield, R. (2011). *SuperCooperators: Altruism, evolution, and why we need each other to succeed*. Simon and Schuster.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131, PubMed: 16251955

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

https://doi.org/10.1016/0028-3932(71)90067-4, PubMed: 5146491

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487. https://doi.org/10.1016/j.conb.2004.07.007, PubMed: 15321069

Paunov, A. M. (2019). *fMRI studies of the relationship between language and theory of mind in adult cognition* [Unpublished doctoral dissertation]. Massachusetts Institute of Technology.

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and theory of mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018, PubMed: 30601693

Paunov, A. M., Shain, C., Chen, X., Lipkin, B., & Fedorenko, E. (2022). *No evidence of theory of mind reasoning in the human language network* [Manuscript in preparation]. Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004, PubMed: 16406760

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167, PubMed: 28053326

Pritchett, B. L., Hoeflin, C., Koldewyn, K., Dechter, E., & Fedorenko, E. (2018). High-level language processing regions are not engaged in action observation or imitation. *Journal of Neurophysiology*, *120*(5), 2555–2570. https://doi.org/10.1152/jn.00222.2018, PubMed: 30156457

Redcay, E., Velnoskey, K. R., & Rowe, M. L. (2016). Perceived communicative intent in gesture and language modulates the superior temporal sulcus. *Human Brain Mapping*, *37*(10), 3444–3461. https://doi.org/10.1002/hbm.23251, PubMed: 27238550

Reddy, A. J., & Wehbe, L. (2020). Syntactic representations in the human brain: Beyond effort-based metrics. *BioRxiv*. https://doi.org/10.1101/2020.06.16.155499

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013, PubMed: 24089502

Reher, K. (Producer), & Sohn, P. (Writer & Director). (2009, May 29). *Partly cloudy* [Animated short film]. Walt Disney Pictures. Pixar Animation Studios.

Richardson, H., Koster-Hale, J., Caselli, N., Magid, R., Benedict, R., Olson, H., Pyers, J., & Saxe, R. (2020). Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *Nature Communications*, *11*, Article 3246. https://doi.org/10.1038/s41467-020-17004-y, PubMed: 32591503

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*, Article 1027. https://doi.org/10.1038/s41467-018-03399-2, PubMed: 29531321

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269. https://doi.org/10.1093/cercor/bhi009, PubMed: 15635062

Ruby, P., & Decety, J. (2003). What you believe versus what you think they believe: A neuroimaging study of conceptual perspective-taking. *European Journal of Neuroscience*, *17*(11), 2475–2480. https://doi.org/10.1046/j.1460-9568.2003.02673.x, PubMed: 12814380

Saur, D., Schelter, B., Schnell, S., Kratochvil, D., Küpper, H., Kellmeyer, P., Klöppel, S., Glaucheac, V., Langeac, R., Mader, W., Feess, D., Timmer, J., & Weiller, C. (2010). Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. *NeuroImage*, *49*(4), 3187–3197. https://doi.org/10.1016/j.neuroimage.2009.11.009, PubMed: 19913624

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062, PubMed: 16635578

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1, PubMed: 12948738

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x, PubMed: 16913952

Saxe, R., Schulz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience*, *1*(3–4), 284–298. https://doi.org/10.1080/17470910601000446, PubMed: 18633794

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development*, *80*(4), 1197–1209. https://doi.org/10.1111/j.1467-8624.2009.01325.x, PubMed: 19630902

Schmälzle, R., Häcker, F. E., Honey, C. J., & Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience*, *10*(8), 1137–1143. https://doi.org/10.1093/scan/nsu168, PubMed: 25653012

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466, PubMed: 27386919

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2021). Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *BioRxiv*. https://doi.org/10.1101/2021.09.18.460917

Shain, C., Blank, I. A., Van Shijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, Article 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307, PubMed: 31874149

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687–E4696. https://doi.org/10.1073/pnas.1323812111, PubMed: 25267658

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's *z* transformation be used? *Journal of Applied Psychology*, *72*(1), 146–148. https://doi.org/10.1037/0021-9010.72.1.146

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society 106*(6), 467–482.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J., Petersson, K. M., & Hagoort, P. (2009). Retrieval and unification of syntactic structure in sentence comprehension: An fMRI study using word-category ambiguity. *Cerebral Cortex*, *19*(7), 1493–1503. https://doi.org/10.1093/cercor/bhn187, PubMed: 19001084

Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1650), 2529–2536. https://doi.org/10.1098/rspb.2008.0762, PubMed: 18664435

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004, PubMed: 31257145

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, *17*(1–2), 3–23. https://doi.org/10.1111/1468-0017.00186

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J. B., & Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, *63*(1), 25–39. https://doi.org/10.1016/j.neuroimage.2012.06.046, PubMed: 22766167

Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and communication in autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 335–364). Wiley. https://doi.org/10.1002/9780470939345

Tahmasebi, A. M., Davis, M. H., Wild, C. J., Rodd, J. M., Hakyemez, H., Abolmaesumi, P., & Johnsrude, I. S. (2012). Is the link between anatomical structure and function equally strong at all cognitive levels of processing? *Cerebral Cortex*, *22*(7), 1593–1603. https://doi.org/10.1093/cercor/bhr205, PubMed: 21893681

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., & Farmer, J. D. (1992). Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, *58*(1–4), 77–94. https://doi.org/10.1016/0167-2789(92)90102-S

Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology*, *44*(3), 187–194. https://doi.org/10.1002/ejsp.2015, PubMed: 25641998

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978, PubMed: 11771995

van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S. A. (2012). Pragmatics in action: Indirect requests engage theory of mind areas and the cortical motor network. *Journal of Cognitive Neuroscience*, *24*(11), 2237–2247. https://doi.org/10.1162/jocn_a_00274, PubMed: 22849399

Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and "theory of mind" in an agrammatic aphasic patient. *Current Biology*, *10*(12), 723–726. https://doi.org/10.1016/s0960-9822(00)00538-8, PubMed: 10873809

Varley, R., Siegal, M., & Want, S. C. (2001). Severe impairment in grammar does not preclude theory of mind. *Neurocase*, *7*(6), 489–493. https://doi.org/10.1093/neucas/7.6.489, PubMed: 11788740

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. R., & Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*(1), 170–181. https://doi.org/10.1006/nimg.2001.0789, PubMed: 11525326

Wehbe, L., Blank, I., Shain, C., Futrell, R., Levy, R., Malsburg, T. Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065, PubMed: 33895807

Whitfield-Gabrieli, S., & Nieto-Castañón, A. (2012). *Conn*: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, *2*(3), 125–141. https://doi.org/10.1089/brain.2012.0073, PubMed: 22642651

Wiesmann, C. G., Schreiber, J., Singer, T., Steinbeis, N., & Friederici, A. D. (2017). White matter maturation is associated with the emergence of theory of mind in early childhood. *Nature Communications*, *8*, Article 14692. https://doi.org/10.1038/ncomms14692, PubMed: 28322222

Wikipedia. (n.d.). *Tree*. https://en.wikipedia.org/wiki/Tree

Wilkinson, K. M. (1998). Profiles of language and communication skills in autism. *Mental Retardation and Developmental Disabilities Research Reviews*, *4*(2), 73–79. https://doi.org/10.1002/(SICI)1098-2779(1998)4:2<73::AID-MRDD3>3.0.CO;2-Y

Willems, R. M., Benn, Y., Hagoort, P., Toni, I., & Varley, R. (2011). Communicating without a functioning language system: Implications for the role of language in mentalizing. *Neuropsychologia*, *49*(11), 3130–3135. https://doi.org/10.1016/j.neuropsychologia.2011.07.023, PubMed: 21810434

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for neuroscience. *Trends in Neurosciences*, *43*(5), 271–273. https://doi.org/10.1016/j.tins.2020.03.003, PubMed: 32353331

Willems, R. M., Van der Haegen, L., Fisher, S. E., & Francks, C. (2014). On the other hand: Including left-handers in cognitive neuroscience and neurogenetics. *Nature Reviews Neuroscience*, *15*(3), 193–201. https://doi.org/10.1038/nrn3679, PubMed: 24518415

Wilson, S. M., Molnar-Szakacs, I., & Iacoboni, M. (2007). Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. *Cerebral Cortex*, *18*(1), 230–242. https://doi.org/10.1093/cercor/bhm049, PubMed: 17504783

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5, PubMed: 6681741

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*(4), 1912–1920. https://doi.org/10.1016/j.neuroimage.2008.01.057, PubMed: 18342544