



Published in final edited form as:

Cortex. 2023 May ; 162: 96–114. doi:10.1016/j.cortex.2023.01.013.

Non-literal language processing is jointly supported by the language and Theory of Mind networks: Evidence from a novel meta-analytic fMRI approach

Miriam Hauptman^{1,2,3}, Idan Blank^{†,1,2,4,5}, Evelina Fedorenko^{†,1,2,6}

¹Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA

²McGovern Institute for Brain Research, MIT, Cambridge, MA 02139, USA

³Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, MD 21218, USA

⁴Department of Psychology, UCLA, Los Angeles, CA 90095, USA

⁵Department of Linguistics, UCLA, Los Angeles, CA 90095, USA

⁶Program in Speech and Hearing in Bioscience and Technology, Harvard University, Boston, MA 02114, USA

Abstract

Going beyond the literal meaning of language is key to communicative success. However, the mechanisms that support non-literal inferences remain debated. Using a novel meta-analytic approach, we evaluate the contribution of linguistic, social-cognitive, and executive mechanisms to non-literal interpretation. We identified 74 fMRI experiments ($n = 1,430$ participants) from 2001–2021 that contrasted non-literal language comprehension with a literal control condition, spanning ten phenomena (e.g., metaphor, irony, indirect speech). Applying the activation likelihood estimation approach to the 825 activation peaks yielded six left-lateralized clusters. We then evaluated the locations of both the individual-study peaks and the clusters against probabilistic functional atlases (cf. anatomical locations, as is typically done) for three candidate brain networks—the language-selective network (Fedorenko et al., 2011), which supports language processing, the Theory of Mind (ToM) network (Saxe & Kanwisher, 2003), which supports social inferences, and the domain-general Multiple-Demand (MD) network (Duncan, 2010), which

Corresponding Authors Miriam Hauptman and Ev Fedorenko, mhauptm1@jhu.edu and evelina9@mit.edu; 43 Vassar Street, Room 46-3037, Cambridge, MA, 02139.

[†]Co-senior authors

Credit author statement

Miriam Hauptman: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Idan Blank:** Conceptualization, Methodology, Supervision, Writing – Review & Editing. **Evelina Fedorenko:** Conceptualization, Methodology, Supervision, Writing – Review & Editing.

Open practices

Materials and data for the study are available at: <https://osf.io/wfsnj/> No part of the study procedures or analyses was pre-registered prior to the research being conducted.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

supports executive control. These atlases were created by overlaying individual activation maps of participants who performed robust and extensively validated ‘localizer’ tasks that selectively target each network in question (n = 806 for language; n = 198 for ToM; n = 691 for MD). We found that both the individual-study peaks and the ALE clusters fell primarily within the language network and the ToM network. These results suggest that non-literal processing is supported by both i) mechanisms that process literal linguistic meaning, and ii) mechanisms that support general social inference. They thus undermine a strong divide between literal and non-literal aspects of language and challenge the claim that non-literal processing requires additional executive resources.

Introduction

Communicative success often requires going beyond accessing literal meanings of words and combining words to construct phrases and sentences (e.g., Grice, 1975; Sperber & Wilson, 1986). Knowledge of ‘non-literal’ meanings of words and phrases (e.g., metaphors, idioms), as well as reliance on contextual information and extra-linguistic cues, like prosody and gestures, are in many cases necessary for apprehending the intended meaning of linguistic input. The neurocognitive mechanisms that support lexical access and phrase-structure building in comprehension and production are fairly well characterized (e.g., Fedorenko et al., 2020), but the mechanisms that enable comprehension beyond those core linguistic processes remain debated. In the remainder of the paper, we use the term ‘non-literal language comprehension’ to refer to the gamut of cognitive processes related to language comprehension that go beyond lexical access and phrase-structure building. Our use of this term is broader than in some other prior papers and encompasses a) classic non-literal phenomena like metaphors; b) discourse-level comprehension; c) pragmatic phenomena like irony; and d) prosody.

Early patient investigations linked difficulties in non-literal interpretation with damage to the right hemisphere (RH) (e.g., Winner & Gardner, 1977; Myers & Linebaugh, 1981; Delis et al., 1983; Brownell et al., 1983; 1986; Bryan, 1988; Joannette et al., 1989; Weylman et al., 1989; Burgess & Chiarello, 1996; Myers, 1998). This work motivated the coarse semantic coding hypothesis (Beeman & Chiarello, 1998; Jung-Beeman, 2005), which posits a RH advantage for processing distantly related concepts, a central aspect of understanding non-literal phenomena such as metaphor. More recently, a growing number of empirical studies have challenged the notion of RH dominance in non-literal processing: brain imaging experiments (e.g., Rapp et al., 2004; Lee & Dapretto, 2006; Bosco et al., 2017; see also Oliveri et al., 2004), meta-analyses of such studies (e.g., Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018), and patient investigations (Ianni et al., 2014; Cardillo et al., 2018; Klooster et al., 2020; see Giora et al., 2000; Zaidel et al., 2002; Klepousniotou & Baum, 2005 for evidence of bilateral involvement) have instead implicated the left hemisphere (LH).

Importantly, both the left and the right hemisphere each contain multiple distinct functional networks, including the language-selective network (e.g., Fedorenko et al., 2011) and its right homotope, the Theory of Mind network (e.g., Saxe & Kanwisher, 2003), and the domain-general Multiple Demand network (e.g., Duncan, 2010). These networks are

associated with distinct cognitive operations, all of which have been argued to contribute to non-literal language comprehension, including in individuals with communication disorders: linguistic processing (e.g., Papagno & Genoni, 2004; Beaty & Silvia, 2013; Whyte & Nelson, 2015), social inference (e.g., Sperber & Wilson, 1986; Happé, 1993; Winner et al., 1998), and executive control (e.g., McDonald & Pearce, 1998; Champagne-Lavau & Stip, 2010), respectively. As a result, focusing on the question of hemispheric dominance alone has limited utility with regard to questions about the cognitive mechanisms underlying non-literal language processing.

In an effort to characterize the cognitive mechanisms underlying non-literal language processing, past neuroimaging studies have primarily relied on the anatomical locations of group-level effects to identify the relevant cognitive processes. However, the traditional group-averaging approach is limited in the inferences it affords about cognitive processes. In particular, in this approach, individual activation maps are averaged in the common space and the resulting activation peaks are interpreted via reverse inference from anatomy to function (Poldrack, 2006, 2011; Fedorenko, 2021). For example, an activation peak in a study on non-literal comprehension that falls within the inferior frontal gyrus may be interpreted as indexing the engagement of cognitive control mechanisms because some past studies that targeted cognitive control reported activation there. Such inferences are not warranted because functional areas vary substantially in their precise locations across individual brains, particularly within the association cortex (e.g., Fischl et al., 2008; Frost & Goebel, 2011; Tahmasebi et al., 2012). Consequently, any given location in the group space may correspond to distinct networks across individual participants (e.g., the language network in one participant, and the Multiple Demand network in another participant; Fedorenko & Blank, 2020).

The use of individual-subject analyses, or ‘precision fMRI’, offers a way to circumvent high inter-individual variability in the locations of functional networks. In this approach, the relevant areas are identified in individual brains using robust and extensively validated ‘localizer’ tasks that selectively target particular functional areas/networks (e.g., Kanwisher et al., 1997; Saxe et al., 2006; Fedorenko et al., 2010, 2013; Shashidara et al., 2019; Fedorenko, 2021; Gratton & Braga, 2021). The recruitment of these areas during some new, critical condition(s) are then examined. For example, a researcher may identify the Multiple Demand network using a robust working-memory-based localizer, and then ask whether these areas, which are active when participants engage in working memory tasks, are also active when participants comprehend a particular type of non-literal language. In this way, the functional localization approach allows for a much more straightforward interpretation of a phenomenon with respect to its underlying cognitive processes and has already helped address many questions that could not be answered using the traditional group-averaging approach (e.g., Fedorenko et al., 2011; Deen et al., 2015; Braga & Buckner, 2017).

Despite the limitations of the group-averaging approach, we would ideally not abandon many hundreds of past fMRI studies. Here we develop an analysis method that integrates individual-subject functional localization and traditional group-averaging experiments and apply it to studies of linguistic phenomena that fall within our broad definition of non-literal language processing. The critical innovation that has enabled this method is the

use of *probabilistic functional atlases* (e.g., Dvoretzky et al., 2021; Thirion et al., 2021; Lipkin et al., 2022). Such atlases are created from large numbers of individuals ($n > 100$ for all atlases used here) who have performed extensively validated functional localizers that selectively target particular functional networks. Because these atlases capture inter-individual variability in the locations of the relevant networks, we can estimate for any given location in the common space the probability that it belongs to each candidate network. In other words, the probabilistic atlases provide a layer of information beyond anatomy.

This new approach differs from other commonly used meta-analytic approaches, particularly the Activation Likelihood Estimation (ALE; Turkeltaub et al., 2002; Eickhoff et al., 2009) approach. The interpretation of ALE estimates still relies on reverse inference from anatomical landmarks, or on comparisons between the results of multiple meta-analyses performed on group-level data. As such, inter-individual variability in functional architecture is not taken into account. A consequence is that the functional resolution of ALE—its ability to discriminate between nearby functional networks—is low (the logic described in Nieto-Castañón & Fedorenko, 2012 for simple group analyses also applies to ALE meta-analyses).

The current approach also offers different information from the information that can be obtained from the NeuroSynth database (Yarkoni et al., 2011). NeuroSynth contains information about activation peaks that are reported in a large number of past group-level fMRI studies. These peaks are associated with keywords that correspond to different cognitive constructs. However, Neurosynth i) extracts activation peaks from the results tables regardless of the nature of the contrast (e.g., for a study that compares a literal and a non-literal condition for some phenomenon, it would extract peaks for both the non-literal>literal contrast and the literal>non-literal contrasts; and ii) extracts keywords regardless of how they are used in a given paper (e.g., for a study that argues that general working memory brain areas are dissociated from brain areas that support language processing, NeuroSynth would extract keywords such as “working memory” and “language”, and whatever peaks are reported in the results tables—distinct peaks for the two sets of areas—would become linked with both keywords). Because the probabilistic atlases we use are constructed using tasks that have been *selectively* and robustly linked to particular cognitive processes, relating the locations of peak activations to these atlases affords both a higher degree of interpretability and a straightforward way to link the results to those from studies that rely on individual-subject functional localization.

In the present study, we examined data from past fMRI studies (74 studies, 825 activation peaks) that contrasted neural responses to non-literal vs. literal conditions across diverse phenomena with respect to three candidate networks: the language-selective network (Fedorenko et al., 2011), which supports language processing, the Theory of Mind (ToM) network (Saxe & Kanwisher, 2003), which supports social inferences, and the domain-general Multiple Demand (MD) network (Duncan, 2010), which supports executive control. In addition to incorporating probabilistic functional atlases in a novel manner, our meta-analysis improves upon past meta-analyses of non-literal language (e.g., Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018) in that we a) include a larger number of studies, b) focus on targeted, more interpretable contrasts (i.e., non-literal > literal; cf. non-literal > fixation), and c) examine a larger number of phenomena that fall within a

broad definition of non-literal language. To foreshadow our results, we find support for the role of the language-selective and ToM networks, but not the MD network, in non-literal interpretation. Further, in line with past meta-analyses of non-literal language processing (Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018), we do not find support for an RH bias: the majority of peaks fall in the LH.

Materials and Methods

In the following sections, we report how we determined our sample size, all inclusion/exclusion criteria (which were established prior to data analysis), all manipulations, and all measures in the study. Materials and data for the study are available at: <https://osf.io/wfsnj/> No part of the study procedures or analyses was pre-registered prior to the research being conducted.

Article selection criteria

A literature search was conducted in accordance with PRISMA guidelines (Moher et al., 2009). Relevant studies were identified in *NeuroSynth*, *Google Scholar*, *APA PsycInfo*, and *PubMed* databases using Boolean searches containing each of the following keywords: “anaphora,” “anthropomorphism,” “comedy,” “discourse comprehension,” “figurative language,” “figure of speech,” “hyperbole,” “humor,” “idioms,” “indirect request,” “indirect speech,” “ironic,” “irony,” “jokes,” “lying,” “metaphor,” “metonymy,” “narrative,” “non-literal language,” “oxymoron,” “paradox,” “personification,” “platitude,” “pragmatics,” “prosody,” “proverbs,” “pun,” “sarcasm,” “sarcastic,” “saying,” “speech act,” “synecdoche,” “text coherence,” “text comprehension,” and “understatement”, plus the disjunctive combination of “fMRI,” “brain,” and “neuroimaging.” All non-literal keywords appeared either in prior fMRI meta-analyses (e.g., Ferstl et al., 2008; Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018) or in theoretical and experimental papers that focus on linguistic phenomena that extend beyond lexical access and phrase-structure building (e.g., Grice, 1975; Demorest et al., 1983; Graesser, Singer, & Trabasso, 1994; Gibbs, 1994; 2002; Colston & O’Brien, 2000).

In addition to conducting database searches using non-literal language keywords, we also examined the reference lists of past neuroimaging meta-analyses on non-literal language processing to minimize the possibility of missing relevant studies (Ferstl et al., 2008; Bohrn et al., 2012; Rapp et al., 2012; Vartanian, 2012; Vrticka et al., 2013; Lisofsky et al., 2014; Yang, 2014; Yang & Shu, 2016; Reyes-Aguilar et al., 2018; Farkas et al., 2021).

260 articles were selected for full-text screening based on the content of their abstracts. The selection criteria included: (1) fMRI (not PET or MEG/EEG) was used; (2) participants were neurotypical, and not aging, adults; (3) participants were native speakers of the language in which the experiment was conducted; (4) a standard whole-brain random effects analysis (Holmes & Friston, 1998) was performed; (5) activation peaks were reported in Talairach (Talairach & Tournoux, 1988) or Montreal Neurological Institute (MNI) (Evans et al., 1993) coordinate systems; and (6) contrasts targeted non-literal language comprehension, in the listening or reading modality, versus a literal (or “less non-literal”) linguistic baseline (cf. coarser-grain contrasts like non-literal language processing vs. fixation). The 74 studies that

satisfied these criteria were published between 2001 and 2021 and targeted ten linguistic phenomena (Table 1). Across these 74 studies, 825 activation peaks (from 102 contrasts, between 1 and 4 contrasts per study) were extracted for analysis. Participants included 1,430 individuals (between 8 and 39 individuals per study; $M = 19.2$) aged 18 to 55 ($M = 24.8$ years), 55% female (see SI Table 1 for further details). Table 1 summarizes the distribution of studies and peaks across the ten phenomena.

The probabilistic functional atlases for the three brain networks of interest

The individual activation peaks and the clusters derived from these peaks via the standard activation likelihood estimation (ALE) analysis (e.g., Turkeltaub et al., 2002; Eickhoff et al., 2009), as described below (section Activation likelihood estimation (ALE) analysis), were evaluated with respect to probabilistic functional atlases for three candidate networks of interest: the language network, the Theory of Mind (ToM) network, and the Multiple Demand (MD) network. For each network, an activation overlap map was created by overlaying a large ($n > 100$) number of individual, binarized activation maps for the ‘localizer’ task targeting that network, as described below (this is the first step in the Group-Constrained Subject-Specific analytic approach, as described in Fedorenko et al., 2010 and Julian et al., 2012). To account for inter-individual variability in the overall level of activation, we selected in each individual the top 10% most localizer-responsive voxels across the brain (fixed-statistical-threshold approaches yield near-identical results; Lipkin et al., 2022). Specifically, we sorted the t -values for the relevant contrast and took 10% of voxels per participant with the highest values. In the resulting activation overlap map, the value in each voxel represents the number of participants for whom that voxel belongs to the top 10% most localizer-responsive voxels. These values—turned into proportions by dividing each value by the total number of participants that contribute to the overlap map—can then be used to estimate the probability that a given voxel belongs to the target network. Consider the extreme cases: if a voxel does not belong to the top 10% most localizer-responsive voxels in *any* participant, that voxel is extremely unlikely to belong to the network of interest, and if a voxel belongs to the top 10% most localizer-responsive voxels in *every* participant, that voxel is extremely likely to belong to the network of interest. In practice, for brain networks that support high-level cognitive functions and fall within the association cortex, network probability is unlikely to ever be 1 because of high inter-individual variability in the precise locations of these networks (e.g., Frost & Goebel, 2011; Tahmasebi et al., 2012), as discussed above. Our degree of confidence in assigning a voxel to a network will therefore be constrained by the maximum inter-subject overlap value for that network.

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, including phonological, lexical-semantic, and combinatorial (semantic and syntactic) processes (e.g., Fedorenko et al., 2010, 2012, 2016, 2020; Bautista & Wilson, 2016; Blank et al., 2016; Regev et al., 2021). It also identifies right-hemisphere homotopes of the left-hemisphere language regions (e.g., Mahowald & Fedorenko, 2016), which have been proposed to play a role in non-literal language comprehension / pragmatic reasoning (e.g., Joannette et al., 1990; Kuperberg et al., 2000; Mashal et al., 2005; Coulson & Williams, 2005; Diaz & Hogstrom,

2011; Eviatar & Just, 2006). Briefly, we used a reading task that contrasted sentences (the critical condition) and lists of unconnected, pronounceable nonwords (the control condition; Figure 1) in a standard blocked design with a counterbalanced condition order across runs. By design, this localizer contrast subtracts out lower-level perceptual (speech or reading-related) and articulatory motor processes (see Fedorenko & Thompson-Schill, 2014 for discussion) and has been shown to generalize across materials, tasks, visual/auditory presentation modality, and languages (e.g., Fedorenko et al., 2010; Fedorenko, 2014; Scott et al., 2017; Ivanova et al., 2020; Chen et al., 2021; Malik-Moraleda, Ayyash et al., 2021). Further, this network emerges robustly from task-free naturalistic data (e.g., Braga et al., 2020; see also Blank et al., 2014, Paunov et al., 2018). Participants read the stimuli one word/nonword at a time in a blocked design, with condition order counterbalanced across runs. Each sentence/nonword sequence was followed by a button-press task to maintain alertness. A version of this localizer is available from <https://evlab.mit.edu/funcloc/download-paradigm>, and the details of the procedure and timing are described in Figure 1 and Table 2. The probabilistic functional atlas used in the current study (Language Atlas (LanA); Lipkin et al., 2022) was constructed using data from 806 participants, and the voxel with the highest network probability had a value of 0.82 (i.e., belonged to the top 10% of most language-responsive voxels in 82% of participants).

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, we used a task based on the classic false belief paradigm (Wimmer & Perner, 1983) that contrasted verbal vignettes about false beliefs (e.g., a protagonist has a false belief about an object's location; the critical condition) versus 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 1). By design, this localizer focuses on ToM reasoning to the exclusion of affective or non-propositional aspects of mentalizing and has been shown to generalize across materials (verbal and non-verbal), tasks, and visual/auditory presentation modality (e.g., Gallagher et al., 2000; Saxe & Kanwisher, 2003; Saxe et al., 2006; Jacoby et al., 2016). Furthermore, this network emerges robustly from task-free naturalistic data (e.g., Braga & Buckner, 2017; DiNicola et al., 2020; see also Paunov et al., 2018). Participants read the vignettes one at a time in a long-event-related design, with condition order counterbalanced across runs. Each vignette was followed by a true/false comprehension question. A version of this localizer is available from <http://saxelab.mit.edu/use-our-efficient-false-belief-localizer>, and the details of the procedure and timing are described in Figure 1 and Table 2. The probabilistic functional atlas used in the current study (unpublished data from the Fedorenko lab) was constructed using data from 198 participants, and the voxel with the highest network probability had a value of 0.88.

The task used to localize the **MD network** is described in detail in Fedorenko et al. (2013) (see also Blank et al., 2014) and targets brain regions that are sensitive to general executive demands. Briefly, we used a spatial working memory task that contrasted a harder and an easier condition. On each trial, participants saw a 3×4 grid and kept track of eight (the critical, harder, condition) or four (the control, easier, condition; Figure 1) locations that were sequentially flashed two at a time or one at a time, respectively. Participants indicated

their memory for these locations in a two-alternative, forced-choice paradigm via a button press. Feedback was provided after every trial. There is ample evidence that demanding tasks of many kinds activate this network (e.g., Duncan & Owen 2000; Fedorenko et al. 2013; Hugdahl et al. 2015; Shashidara et al., 2019; and Assem et al., 2020a). Furthermore, this network emerges robustly from task-free naturalistic data (e.g., Assem et al., 2020a; Braga et al., 2020; see also Blank et al., 2014, Paunov et al., 2019). Hard and easy conditions were presented in a blocked design, with condition order counterbalanced across runs. This localizer is available from the authors upon request and the details of the procedure and timing are described in Figure 1 and Table 2. The probabilistic functional atlas used in the current study (MDAtlas; Lipkin et al., in prep.) was constructed using data from 691 participants, and the voxel with the highest network probability had a value of 0.75. (It is worth noting that the three networks of interest differ somewhat in the range of their non-zero network probability values: language = 0.001–0.82, ToM = 0.005–0.88, and MD = 0.001–0.75. We decided against normalizing these values for each map (so that the highest network probability would be set to 1, or to the highest value observed across any of the three networks) because doing so would obscure meaningful differences in the likelihood that a given voxel belongs to each network.)

Critically, these three networks are robustly spatially and functionally dissociable within individuals in both naturalistic (e.g., Blank et al., 2014; Paunov et al., 2018; Braga et al., 2020) and task-based fMRI paradigms despite their close proximity to each other within the association cortex. The *language network* selectively supports linguistic processing, showing little or no response to diverse executive function tasks (e.g., Fedorenko et al., 2011; Monti et al., 2012; see Fedorenko & Blank, 2020 for a review) and mentalizing tasks in individual participants (Deen et al., 2015, Paunov, 2018; Paunov et al., 2021; Shain, Paunov, Chen et al., 2022). The data from the patient literature mirrors this selectivity: damage to the language network does not appear to lead to difficulties in executive or social processing (see Fedorenko and Varley, 2016 for a review). The *ToM network* selectively supports social cognition, showing little or no response to linguistic input without mental state content (Deen et al., 2015; Paunov et al., 2021; Deen & Freiwald, 2021; Shain, Paunov, Chen et al., 2022) or to executive demands (Saxe et al., 2006; Scholz et al., 2009; Willems et al., 2010). Finally, the *MD network* supports diverse executive demands (e.g., Duncan, 2010, 2013; Assem et al., 2020a; Smith et al., 2021) and is linked to fluid reasoning ability (e.g., Woolgar et al., 2010; Assem et al., 2020b), but plays a limited role in language comprehension once task demands are controlled for (e.g., Diachek, Blank, Siegelman et al., 2020; Shain, Blank et al., 2020; Wehbe et al., 2021; see Fedorenko & Shain, 2021 for a review), and in social cognition (e.g., Willems et al., 2010). Because of the selective relationship between each of the three networks and a particular set of cognitive processes, activity in these networks can be used as an index of the engagement of the relevant processes (see e.g., Mather et al., 2013, for discussion), circumventing the need for precarious reverse inference from anatomical locations to function (e.g., Poldrack et al., 2006, 2011; Fedorenko, 2021). This is the approach that is adopted in studies that rely on functional localization (e.g., Brett et al., 2002; Saxe et al., 2006; Fedorenko, 2021). Here, we extend this general logic to the meta-analysis of group-level activation peaks by leveraging

information about the landscape of each network of interest based on probabilistic functional atlases for the relevant localizer tasks.

Extracting network probabilities for the individual-study peaks

Prior to analysis, activation peaks that were reported in Talairach space were converted to the MNI space using `icbm2tal` transform SPM conversion in GingerALE 3.0.2 (Eickhoff et al., 2009, 2012). The MNI coordinates of the 825 activation peaks included in the dataset were tested against the probabilistic functional atlases (created as described above, section The probabilistic functional atlases for the three brain networks of interest) for each of the three functional networks of interest. For each coordinate, three network probability values were extracted (one from each functional atlas). (Note that because of the variability in the precise locations of functional areas and the proximity of the three networks to each other in parts of the association cortex, many voxels have non-zero values for more than a single network (despite the fact that there is little to no overlap among the networks in individual participants)—this is precisely the argument against traditional group-averaging analyses in fMRI; e.g., Fedorenko & Blank, 2020; DiNicola & Buckner, 2021; Fedorenko, 2021; Gordon & Nelson, 2021; Gratton & Braga, 2021; Smith et al., 2021). It is also important to note that what we refer to throughout the manuscript as ‘probability values’ (which may be taken to mean posterior probabilities, i.e., $p(\text{language} \mid \text{voxel})$) are, in fact, likelihood values (e.g., $p(\text{voxel} \mid \text{language})$, i.e., the probability that a voxel will be active during linguistic processing based on the percentage of people who show activity in that voxel for the language localizer contrast). This distinction does not matter given that our critical tests compare *across* different networks (rather than, for example, trying to estimate an absolute belief for any particular network).

Activation likelihood estimation (ALE) analysis

In addition to examining the individual-study activation peaks, we used the GingerALE software (Eickhoff et al., 2009, 2012) to perform a traditional fMRI meta-analysis via the activation likelihood estimation (ALE) method (Turkeltaub et al., 2002). ALE identifies regions that show consistent activation across experiments (Eickhoff et al., 2009, 2012). The clusters that ALE yields should be less noisy than the individual-study activation peaks, especially when multiple-comparison correction is not appropriately applied and participant sample sizes are small in individual studies (e.g., Genovese et al., 2002; Eklund et al., 2016; Chen et al., 2018).

The 74 studies in our dataset yielded 825 activation peaks (Table 1; see SI Table 1 for a complete list of studies), but 39 peaks (fewer than 5%) fell outside the MNI template used by GingerALE and were therefore excluded, leaving 786 peaks. For each study, a map was created in which each voxel in the MNI space received a modeled activation score. Modeled activation scores reflect the likelihood that significant activation for one particular experiment was observed at a given voxel. The 74 modeled activation maps were then unified, generating an ALE value for each voxel.

Significance was assessed by comparing the observed ALE values to a null distribution that was generated by repeatedly calculating ALE values using randomly placed activation

peaks (1,000 permutations). A cluster-forming threshold of $p < 0.001$ (uncorrected) identified contiguous volumes of significant voxels (“clusters”), and clusters that survived a cluster-level family-wise error (FWE) of $p < 0.05$ were considered significant. Cluster-level FWE correction has been argued to be the most appropriate correction for ALE, as it minimizes false positives while remaining more sensitive to true effects in comparison to other correction methods (Eickhoff et al., 2016).

Applying ALE to the 74 experiments yielded six significant clusters that were located in the left frontal and temporal cortex, and in the left amygdala (SI Figure 1). The clusters varied in size from 176 to 1,695 voxels.

Critical analyses

We asked two key research questions: the first, critical question examines the locations of the activation peaks with respect to the *three brain networks of interest*: the language network, the ToM network, and the MD network, and the second question is motivated by the prior claim about the privileged role of the *right hemisphere* in non-literal language processing (e.g., Winner & Gardner, 1977). Note that in all the analyses, we collapse the data from across the ten non-literal phenomena. We adopted this approach because i) none of the phenomena have a sufficient number of studies to enable meaningful phenomenon-level examination, and ii) there is currently a lack of consensus about the ways to carve up the space of non-literal phenomena (see Discussion). However, in an exploratory analysis, we examined the contribution of different phenomena to the peaks/clusters that load on the functional networks.

Q1: How are the activation peaks from prior fMRI studies of non-literal language comprehension distributed across the language, ToM, and MD networks?

Analysis of individual-study peaks: We evaluated the locations of the activation peaks with respect to our three brain networks of interest (language, ToM, MD). Almost all the peaks—820/825 (99%)—were observed in voxels with a non-zero network probability in at least one functional atlas. To test whether the network probabilities associated with the activation peaks differed between the networks, we developed the following statistical procedure (note that for all analyses, we excluded the 13 peaks that fell on the cortical midline, i.e., had an x coordinate value of 0, given that we wanted to examine each network in each hemisphere separately, which left 812 peaks for analysis):

1. For each of the 812 peaks, we calculated the difference in network probabilities between each pair of networks (language vs. ToM, language vs. MD, and ToM vs. MD), and recorded the median (across peaks) of these difference values for each of the three contrasts in each hemisphere separately. (We used median instead of mean values to better capture the central tendencies given the skewness of the distributions.)
2. We then created 3D maps from which random peak sets could be selected (against which the location of true observed peaks could be evaluated, as described in Step 4 below). Specifically, we created a map for each hemisphere

whereby we removed the voxels whose locations were associated with a network probability of 0 in all three functional atlases (for the language, ToM, and MD networks). This restriction was imposed so as to a) constrain the locations of the baseline voxel sets to the parts of the brain where true observed peaks were found (as noted above, almost all the peaks were observed in voxels that had a non-zero network probability in at least one network); and thus, b) to construct a more conservative test, making it more difficult to detect between-network differences.

3. In each hemisphere, the same number of peaks as in our dataset ($n = 490$ in the LH, and $n = 322$ in the RH) were then randomly sampled from the maps that were created in Step 2, and the median difference (across peaks) for each contrast (language vs. ToM, language vs. MD, ToM vs. MD) was computed, as in Step 1. This procedure was repeated 10,000 times, yielding an empirical null distribution of median network probability differences.
4. Finally, we compared the true median network probability differences against the distribution of network probability differences obtained in Step 3 to yield a significance value for each of the three contrasts within each hemisphere.

Analysis of activation likelihood estimation (ALE) clusters: We evaluated the locations of 1) the center peaks of the six significant ALE clusters, and 2) all voxels contained within each cluster with respect to our three brain networks of interest (language, ToM, MD). Importantly, whereas individual study peaks might be skewed by a single study with a particularly large number of activation peaks, ALE estimates incorporate the sample size of each study and are therefore less susceptible to this source of bias. To test whether the network probabilities differed between the networks, we evaluated the network probabilities of the voxels in each cluster and, for each cluster individually, subsequently performed the statistical procedure described above (in Analysis of individual-study peaks).

Q2: Do the activation peaks from prior fMRI studies of non-literal language comprehension exhibit a right hemisphere bias?

Analysis of individual-study peaks: To test whether the two hemispheres differed in the number of the activation peaks, we ran two logistic mixed effect regression models using the “lme4” package in R. The first tested the full set of 812 activation peaks (excluding the 13 peaks that fell on the cortical midline), and the second focused on the subset of the 812 peaks ($n = 636$) that exhibited a network probability of 0.10 or greater in at least one network. The second model was included to ensure that the results are not driven by a subset of peaks that do not load strongly on any of the three networks of interest. For both models, the following formula was used, which included a random intercept for experiment ($n = 74$):

$$\text{Location of a peak in LH or RH} \sim 1 + (1 \mid \text{Experiment})$$

Analysis of activation likelihood estimation (ALE) clusters: We examined the locations of the six significant ALE clusters with respect to hemisphere.

Results

Non-literal language comprehension draws primarily on the language and ToM networks

Across the 812 non-midline peaks, the highest network probability was observed for the ToM network functional atlas (maximum=0.86, mean=0.19, SD=0.18, median=0.12; 802 total peaks with nonzero network probabilities), followed by the language atlas (maximum=0.80, mean=0.17, SD=0.17, median=0.10; 807 nonzero peaks), and the MD atlas (maximum=0.65, mean=0.11, SD=0.14, median=0.04; 806 non-zero peaks). In both hemispheres, the median nonzero network probabilities across the activation peaks numerically exceeded the median non-zero probability values across the functional atlases as a whole: language atlas (LH: peaks=0.12, atlas=0.04; RH: peaks=0.07, atlas=0.04), ToM (LH: peaks=0.13, atlas=0.04; RH: peaks=0.13, atlas=0.04), and MD (LH: peaks=0.03, atlas=0.02; RH: peaks=0.04, atlas=0.02). See Figure 2 for a depiction of the network probability distributions of the individual-study peaks for each network.

We then examined between-network differences in network probabilities. Among the 490 LH peaks, the magnitude of the median difference in network probabilities was highest between the language and MD networks (language vs. MD = 0.07, language vs. ToM = 0.02, ToM vs. MD = 0.04). In the RH (322 peaks), the magnitude of the median difference in network probabilities was highest between the ToM and MD networks (ToM vs. MD = 0.04, language vs. ToM = 0.02, language vs. MD = 0.02). Our permutation analysis (see Methods) revealed that the LH activation peaks were located more centrally in the language network than in either the ToM or the MD network (both p s < 0.0001) and more centrally in the ToM network than the MD network (p < 0.0001). The RH activation peaks were located more centrally in the ToM network than either the language or the MD network (both p s < 0.0001) and more centrally in the language network than the MD network (p < 0.0001) (see Figure 2). Importantly, this analysis controls for network lateralization (language=LH, ToM=RH) by generating an empirical null distribution of probability differences in each hemisphere separately. Higher network probabilities observed in the language network among the LH voxels and in the ToM network among the RH voxels therefore exceed what would be expected given the (already-skewed) distribution of network probabilities in each hemisphere.

With respect to the ALE clusters (see Table 3 and SI Figure 1), the center peaks of four clusters, including the largest cluster, exhibited the highest network probabilities in the language atlas. Two of these were located in the left temporal lobe (primarily within the superior and middle temporal gyri), one within the left inferior frontal gyrus, and one within the left amygdala (although note that the network probability for the amygdala peak is overall relatively low compared to the cortical peaks). The center peaks of two additional clusters exhibited the highest network probabilities in the ToM atlas. One of these clusters was located in the left medial frontal gyrus and one in the left superior and middle temporal gyri.

Analysis of the network probabilities associated with the full set of voxels comprising each ALE cluster yielded similar results (Table 3). The voxels in the four clusters whose center peaks had the highest network probabilities in the language atlas also exhibited the greatest

median network probabilities in the language atlas, whereas the voxels in the two clusters whose center peaks had the highest network probabilities in the ToM atlas also exhibited the greatest *median network probabilities* in the ToM atlas. Results from the permutation analysis supported this apparent distinction between the four “language” clusters and the two “ToM” clusters: voxels comprising the language clusters had significantly higher network probabilities in the language network than in the ToM and MD networks, and voxels comprising the ToM clusters had significantly higher network probabilities in the ToM network than in the language and MD networks (all p s < 0.0001). These results are displayed in Figure 3. In general, the median probability values obtained from the voxels comprising ALE clusters are numerically higher than those observed in the individual activation peaks that were submitted to the same analysis. This likely reflects the inherent noisiness of individual activation peaks culled from experiments that use traditional group analyses and varied statistical correction approaches. However, the probability values associated with the individual activation peaks in our dataset, although numerically low, are not random (see SI Figure 2 and SI Figure 3).

Finally, in an exploratory analysis, we evaluated the phenomena that contributed to the six significant ALE clusters (see SI Table 2). No clear differentiation in terms of phenomena that contribute to the language vs. the ToM clusters was apparent.

No evidence of an RH bias for non-literal language comprehension

Of the full set of 812 peaks (excluding the 13 that fell on the cortical midline), 322 peaks (~40%) were located in the RH, and 490 (~60%) were located in the LH (Figure 4). This difference was highly reliable ($b = -0.44$, $SE = 0.1$, $z = -4.42$, $p < 0.001$). Similarly, of the subset of the 636 peaks with network probabilities of at least 0.10 in at least one network of interest, 245 peaks (~39%) were located in the RH, and 391 (~61%) were located in the LH (Figure 4). This difference was highly reliable ($b = -0.50$, $SE = 0.11$, $z = -4.5$, $p < 0.001$). ALE analysis yielded six LH clusters and zero RH clusters, suggesting that the RH peaks were less reliable across studies and may therefore reflect spurious results (see also Bohrn et al., 2012), or that RH involvement is specific to a small subset of non-literal phenomena. Collectively, these findings provide additional support for the notion that the LH is an important contributor—perhaps more important than the RH—to non-literal language processing (Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018).

Discussion

To illuminate the cognitive and neural bases of non-literal language comprehension, we performed a meta-analysis of group-level activation peaks from past fMRI studies. Specifically, we developed a novel approach that leverages ‘probabilistic functional atlases’ for three brain networks (the language network, the Theory of Mind network, and the Multiple Demand network) that have been implicated in non-literal language comprehension, broadly construed. The atlases are built using large numbers of individual activation maps for extensively validated ‘localizer’ tasks (e.g., Saxe et al., 2006; Fedorenko, 2021) and provide estimates of the probability that a location in the common brain space belongs to a particular network. Because each of these networks has been rigorously

characterized and selectively linked to particular cognitive processes in past work, activation peak locations therein can be interpreted as evidence for the engagement of the relevant process(es) (Mather et al., 2013). This approach is therefore superior to the traditional meta-analytic approach where activation peaks or ALE clusters (Turkeltaub et al., 2002) are interpreted solely based on their anatomical locations or in relation to the results of additional meta-analyses.

The three networks that we examined include i) the language-selective network, which supports literal comprehension, including lexical and combinatorial operations (Fedorenko et al., 2020), ii) the Theory of Mind (ToM) network, which supports social inference, including mentalizing (Saxe & Kanwisher, 2003), and iii) the Multiple Demand (MD) network, which supports executive control (Duncan, 2010). We asked two research questions. The first, critical question concerned the distribution of peaks across the networks. The individual peaks and ALE clusters tended to fall in the language and ToM networks, but not the MD network. The second question, motivated by past patient investigations that have linked non-literal comprehension impairments to RH damage (e.g., Winner & Gardner, 1977), asked whether the peaks were more likely to fall in the RH. In line with past meta-analyses of non-literal language (Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018), we found that the peaks/clusters fell primarily in the LH. Below, we discuss several issues that these results inform, and highlight some outstanding questions and some methodological implications.

The role of the ToM and language networks in non-literal comprehension

In line with past studies that have reported activations in putative ToM areas for non-literal phenomena (e.g., Spotorno et al., 2012; van Ackeren et al., 2012; Feng et al., 2017), some individual study activation peaks and two ALE clusters had a high probability of falling within the ToM network. What is the role of mentalizing—a capacity supported by the ToM network—in language comprehension?

Because linguistic inputs often underspecify intended meaning (Wittgenstein, 1953; Sperber & Wilson, 1986), language comprehension routinely requires inferences about communicative intent. The computation of such inferences has historically been a focus of the field of pragmatics (Grice, 1957, 1975). Some early proposals drew a sharp boundary between literal and non-literal/pragmatic processing (Grice, 1975; Searle, 1979). However, defining the scope of pragmatic inference has proven challenging, and many have questioned the divide between literal and inferred meaning, or between semantics and pragmatics (Jackendoff, 2002). But if no such boundary exists, does understanding language *always* recruit the ToM network, in addition to the language network, to enable inferences about communicative intent?

The notion of continuous ToM engagement during language comprehension does not seem *a priori* plausible (many phenomena requiring context-based inferences—lexical disambiguation or pronoun resolution—are so common that it would seem inefficient and unnecessary to constantly engage in full-blown mentalizing) and does not find empirical support (Deen et al., 2015; Paunov et al., 2021; Shain, Paunov, Chen et al., 2022). Yet, in some cases, the ToM network does appear to contribute to language comprehension

(e.g., Spotorno et al., 2012). Delineating the precise conditions under which comprehension requires ToM resources remains an important goal for future work (Paunov et al., 2021). Brain-imaging investigations of diverse non-literal phenomena using approaches with functionally localized ToM and language networks may offer some clarity. These approaches could be complemented by experiments that test linguistic abilities in individuals with impaired ToM reasoning (e.g., Siegal et al., 1996; Happé et al., 1999) or in statistical language models (Devlin et al., 2019). The latter can reveal which non-literal phenomena can be handled by language models, and which might require additional machinery that approximates mental state inference (e.g., Hu et al., 2021; 2022).

Many individual study activation peaks and four ALE clusters had a high probability of falling within the language network. Of course, understanding both literal and non-literal language requires language processing, so both conditions should elicit a strong response in the language network. But why do non-literal conditions often elicit a stronger response? Differences in linguistic complexity might provide one explanation. The brain's language areas are sensitive to comprehension difficulty (Wehbe et al., 2021). Although studies that compare literal and non-literal conditions commonly match stimuli on some linguistic variables, this matching is often limited to word-level features, which do not account for potential differences in context-based processing difficulty (e.g., surprisal; Smith & Levy, 2011). Further, despite the robust dissociation between the language and ToM networks, Paunov et al. (2019) found that the two networks show reliable correlation during language processing. This inter-network synchronization may lead increased ToM demands to additionally manifest in the language network via inter-network connections. Finally, the language network may actually support some pragmatic computations, although to argue for such effects, linguistic confounds and inter-network information leakage would need to be eliminated as possible explanations.

No evidence for the role of the MD network in non-literal comprehension

In contrast to the ToM and language networks, the Multiple Demand network does not appear to support non-literal comprehension: individual-study activation peaks and ALE clusters were least likely to fall in this network. Selecting the non-literal interpretation of linguistic input may tax working memory (because multiple interpretations may be activated) or require inhibitory control (e.g., Gernsbacher & Robertson, 1999; Channon & Watts, 2003). However, recent work has shown that any such operation related to linguistic processing appears to be implemented within the language-selective network (Shain, Blank et al., 2020; see Fedorenko & Shain, 2021 for a review). Indeed, as discussed above, greater cognitive demands associated with non-literal interpretation may well explain the responses in the *language* network.

We suspect that neural activity observed during non-literal processing in putative executive control areas in past studies (e.g., AbdulSabur et al., 2014; Chan & Lavalée, 2015; Bosco et al., 2017) is due either to i) reliance on anatomical landmarks, which do not warrant functional interpretation in the association cortex (Poldrack, 2006; Fedorenko, 2021), or ii) extraneous task demands (see Diachek, Blank, Siegelman et al., 2020 for evidence that sentence comprehension only engages the MD network when accompanied by a secondary

task, like a sentence judgment). Behavioral investigations that find correlations between executive abilities and non-literal interpretation abilities (e.g., Akbar et al., 2013; Caillies et al., 2014; Rints et al., 2015) are also likely affected by methodological issues, from small sample sizes to confounded experimental paradigms (Matthews et al., 2018). Indeed, recent large-scale investigations (Cardillo et al., 2020; Fairchild & Papafragou, 2021) argue against the role of executive functions in pragmatic ability.

Possible reasons for the inconsistencies regarding the hemispheric bias

One remaining puzzle concerns the difference between patient work, which has implicated the RH in non-literal language processing (e.g., Winner & Gardner, 1977; Myers & Linebaugh, 1981; Delis et al., 1983; Brownell et al., 1983; 1986; Van Lancker & Kempler, 1987; Brownell, 1988; Stemmer et al., 1994; Giora et al., 2000; Ferstl et al., 2005; see also Jung-Beeman, 2005) and our results, which demonstrate a LH bias. As noted above, left-lateralized activations have also been reported in past studies (e.g., Rapp et al., 2004; Lee & Dapretto, 2006; Hillert & Bura as, 2009; Piñango et al., 2015; Bosco et al., 2017) and meta-analyses (Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018; Farkas et al., 2021). One possibility is that no RH bias exists for non-literal language processing. Both hemispheres contribute, perhaps with the LH contributing more strongly, as per the standard LH language bias (e.g., Geschwind, 1970). Because lexical and grammatical impairments resulting from LH damage are more salient and devastating, non-literal comprehension difficulties go unnoticed. In contrast, RH damage, which does not strongly affect basic language processing, may make apparent more subtle linguistic impairments, leading to the apparent RH bias for non-literal comprehension. In line with this possibility, several studies have reported that patients with LH damage show similar, or even greater, deficits in non-literal comprehension compared to RH-damaged patients (e.g., Tompkins, 1990; Giora et al., 2000; Zaidel et al., 2002; Gagnon et al., 2003; Klepousniotou & Baum, 2005; Ianni et al., 2014; Cardillo et al., 2018; Klooster et al., 2020).

Alternatively, the RH may indeed contribute more strongly than the LH to non-literal language comprehension, and past fMRI studies may have failed to detect the RH effects due the generally low sensitivity of the group-averaging analytic approach (Nieto-Castañón & Fedorenko, 2012). This possibility seems unlikely: if the RH were more active than the LH during non-literal comprehension, then RH activations should be easier to detect in a group analysis. This is because stronger responses would be more spatially extensive in individual brains and thus more likely to lead to overlap at the group level. (Further, using our probabilistic functional atlases, we did not find evidence for the possibility that functional areas in the RH are more variable in their locations across individuals and therefore less likely to emerge in a group analysis.) Future fMRI studies where areas of interest are identified via functional localizers, as well as intracranial stimulation studies, which enable spatially precise perturbations of neural activity, can provide important insight into the hemisphere-bias question.

Comparison with prior meta-analyses of non-literal language

Our results are generally consistent with several prior meta-analyses of non-literal language processing that report activation in left inferior frontal and anterior/middle temporal lobes as

individuals engage in non-literal interpretation (Ferstl et al., 2008; Bohrn et al., 2012; Rapp et al., 2012; Reyes-Aguilar et al., 2018; Rapp, 2019). However, in contrast to previous meta-analyses, which focused on localizing regions that support non-literal language processing (i.e., the “where” question), our approach enables us to make stronger inferences about the underlying cognitive processes—linguistic processing, mentalizing, or executive control—that non-literal comprehension involves (i.e., the “how” question). Collectively, this and past work point to the primacy of linguistic resources during non-literal language comprehension, where “non-literal language” encompasses linguistic phenomena that extend beyond lexical access and phrase-structure building. One possible explanation for the recruitment of left-lateralized language areas is that non-literal language processing places increased demand on accessing meanings of words and constructions and combining them into compositional representations. Activity in the language brain areas reliably scales with the difficulty of language processing; in particular, the magnitude of neural response during naturalistic comprehension strongly relates to how long a given word takes to process, as measured behaviorally (e.g., Wehbe et al., 2021; see also Shain, Blank et al., 2020 and Shain et al., 2022). Given that many non-literal phenomena involve accessing words in unusual senses and/or combining words in novel ways, it is perhaps expected that processing non-literal language would be more linguistically demanding.

Another possibility, informed by recent work on inter-network connections between the language and ToM networks (Paunov et al., 2019, 2022), is that non-literal interpretation elicits increased crosstalk between these networks. The recruitment of ToM areas observed in the present study diverges from the results of prior meta-analyses, where activation in classic ToM regions is less consistently observed, perhaps due to the fact that fewer studies and non-literal phenomena were incorporated. Although our results cannot speak to the existence of a neurally separable pragmatics substrate (see Bendtz et al., 2022), the upregulation of inter-network connections between language and ToM may be unique to non-literal processing; however, as mentioned above, a rigorously controlled set of stimuli would be required to properly address this question (Shain et al., 2022).

The need to carve non-literal processing at its joints

One important limitation of the present study is that there were not a sufficient number of studies to evaluate differences between phenomena (or within phenomena; e.g., novel vs. conventional metaphors). In an exploratory analysis, we did not observe any systematic patterns with regard to the phenomena that preferentially recruit the language vs. the ToM network (SI Table 2). In general, the field would benefit from clear, formal hypotheses about differences in the cognitive processes that contribute to each non-literal phenomenon. Such hypotheses can then be tested using behavioral, brain imaging, and computational modeling approaches.

We would like to conclude with a methodological point. fMRI has been a critical tool in human cognitive neuroscience. However, two disparate approaches are currently in use: i) the traditional approach where brains are averaged voxel-wise in the common space (Friston et al., 1999), and ii) the subject-specific functional localization approach where regions of interest are identified using functional ‘localizers’ (or other precision mapping

approaches, e.g., Braga et al., 2020; DiNicola et al., 2020) and critical effects are examined therein (Saxe et al., 2006). Until recently, comparing findings across these approaches—and thus establishing a cumulative research enterprise—has proven challenging. Probabilistic functional atlases, like the ones we used in the current study (see also Dworetzky et al., 2021; Thirion et al., 2021), can bridge these two approaches by providing a common framework for functional areas/networks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Alex Paunov, Josef Affourtit, and Ben Lipkin for help with some of the analyses, and Ariel Goldberg and Jayden Ziegler for comments on the undergraduate thesis that resulted from this work. This work was supported by the NIH R01 award DC016607 and funds from MIT's Simons Center for the Social Brain (via SFARI); EF was further supported by the NIH R01 award DC016950, and funds from the McGovern Institute for Brain Research and the Brain and Cognitive Sciences department.

References

- AbdulSabur NY, Xu Y, Liu S, Chow HM, Baxter M, Carson J, & Braun AR (2014). Neural correlates and network connectivity underlying narrative production and comprehension: A combined fMRI and PET study. *Cortex*, 57, 107–127. 10.1016/j.cortex.2014.01.017 [PubMed: 24845161]
- Adamczyk P, Wyczesany M, Domagalik A, Daren A, Cepuch K, Bł dzi ski P, Cechnicki A, & Marek T (2017). Neural circuit of verbal humor comprehension in schizophrenia—An fMRI study. *NeuroImage: Clinical*, 15, 525–540. 10.1016/j.nicl.2017.06.005 [PubMed: 28652967]
- Ahrens K, Liu H-L, Lee C-Y, Gong S-P, Fang S-Y, & Hsu Y-Y (2007). Functional MRI of conventional and anomalous metaphors in Mandarin Chinese. *Brain and Language*, 100(2), 163–171. 10.1016/j.bandl.2005.10.004 [PubMed: 16298426]
- Akbar M, Loomis R, & Paul R (2013). The interplay of language on executive functions in children with ASD. *Research in Autism Spectrum Disorders*, 7(3), 494–501. 10.1016/j.rasd.2012.09.001
- Akimoto Y, Sugiura M, Yomogida Y, Miyauchi CM, Miyazawa S, & Kawashima R (2014). Irony comprehension: Social conceptual knowledge and emotional response. *Human Brain Mapping*, 35(4), 1167–1178. 10.1002/hbm.22242 [PubMed: 23408440]
- Assem M, Blank IA, Mineroff Z, Ademo lu A, & Fedorenko E (2020). Activity in the fronto-parietal multiple-demand network is robustly associated with individual differences in working memory and fluid intelligence. *Cortex*, 131, 1–16. 10.1016/j.cortex.2020.06.013 [PubMed: 32777623]
- Assem M, Glasser MF, Van Essen DC, & Duncan J (2020). A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cerebral Cortex (New York, NY)*, 30(8), 4361–4380. 10.1093/cercor/bhaa023
- Ayyash D, Malik-Moraleda S, Gallée J, Affourtit J, Hoffman M, Mineroff Z, Jouravlev O, & Fedorenko E (2021). The universal language network: A cross-linguistic investigation spanning 45 languages and 11 language families (p. 2021.07.28.454040). 10.1101/2021.07.28.454040
- Baggio G (2018). *Meaning in the Brain*. MIT Press.
- Bambini V, Gentili C, Ricciardi E, Bertinetto PM, & Pietrini P (2011). Decomposing metaphor processing at the cognitive and neural level through functional magnetic resonance imaging. *Brain Research Bulletin*, 86(3), 203–216. 10.1016/j.brainresbull.2011.07.015 [PubMed: 21803125]
- Bašnáková J, van Berkum J, Weber K, & Hagoort P (2015). A job interview in the MRI scanner: How does indirectness affect addressees and overhearers? *Neuropsychologia*, 76, 79–91. 10.1016/j.neuropsychologia.2015.03.030 [PubMed: 25858603]

- Bašnáková J, Weber K, Petersson KM, van Berkum J, & Hagoort P (2014). Beyond the Language Given: The Neural Correlates of Inferring Speaker Meaning. *Cerebral Cortex*, 24(10), 2572–2578. 10.1093/cercor/bht112 [PubMed: 23645715]
- Bautista A, & Wilson SM (2016). Neural responses to grammatically and lexically degraded speech. *Language, Cognition and Neuroscience*, 31(4), 567–574. 10.1080/23273798.2015.1123281 [PubMed: 27525290]
- Beaty RE, & Silvia PJ (2013). Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory & Cognition*, 41(2), 255–267. 10.3758/s13421-012-0258-5 [PubMed: 23055118]
- Beeman MJ, & Chiarello C (1998). Complementary Right- and Left-Hemisphere Language Comprehension. *Current Directions in Psychological Science*, 7(1), 2–8. 10.1111/1467-8721.ep11521805
- Bekinschtein TA, Davis MH, Rodd JM, & Owen AM (2011). Why Clowns Taste Funny: The Relationship between Humor and Semantic Ambiguity. *Journal of Neuroscience*, 31(26), 9665–9671. 10.1523/JNEUROSCI.5058-10.2011 [PubMed: 21715632]
- Bendersky M, Lomlomdjian C, Abusamra V, Elizalde Acevedo B, Kochen S, & Alba-Ferrara L (2021). Functional anatomy of idiomatic expressions. *Brain Topography*, 34(4), 489–503. 10.1007/s10548-021-00843-3 [PubMed: 33948754]
- Bendtz K, Ericsson S, Schneider J, Borg J, Bašnáková J, & Uddén J (2022). Individual Differences in Indirect Speech Act Processing Found Outside the Language Network. *Neurobiology of Language*, 3(2), 287–317. 10.1162/nol_a_00066 [PubMed: 37215561]
- Blank I, Balewski Z, Mahowald K, & Fedorenko E (2016). Syntactic processing is distributed across the language system. *NeuroImage*, 127, 307–323. 10.1016/j.neuroimage.2015.11.069 [PubMed: 26666896]
- Blank I, Kanwisher N, & Fedorenko E (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, 112(5), 1105–1118. 10.1152/jn.00884.2013 [PubMed: 24872535]
- Bohrn IC, Altmann U, & Jacobs AM (2012). Looking at the brains behind figurative language —A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia*, 50(11), 2669–2683. 10.1016/j.neuropsychologia.2012.07.021 [PubMed: 22824234]
- Bosco FM, Parola A, Valentini MC, & Morese R (2017). Neural correlates underlying the comprehension of deceitful and ironic communicative intentions. *Cortex*, 94, 73–86. 10.1016/j.cortex.2017.06.010 [PubMed: 28728080]
- Braga RM, & Buckner RL (2017). Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. *Neuron*, 95(2), 457–471.e5. 10.1016/j.neuron.2017.06.038 [PubMed: 28728026]
- Braga RM, DiNicola LM, Becker HC, & Buckner RL (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. 10.1152/jn.00753.2019 [PubMed: 32965153]
- Brett M, Johnsrude IS, & Owen AM (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, 3(3), 243–249. 10.1038/nrn756 [PubMed: 11994756]
- Brownell HH (1988). The neuropsychology of narrative comprehension. *Aphasiology*, 2(3–4), 247–250. 10.1080/02687038808248918
- Brownell HH, Michel D, Powelson J, & Gardner H (1983). Surprise but not coherence: Sensitivity to verbal humor in right-hemisphere patients. *Brain and Language*, 18(1), 20–27. 10.1016/0093-934X(83)90002-0 [PubMed: 6839130]
- Brownell HH, Potter HH, Bihrlle AM, & Gardner H (1986). Inference deficits in right brain-damaged patients. *Brain and Language*, 27(2), 310–321. 10.1016/0093-934X(86)90022-2 [PubMed: 3955344]
- Bryan KL (1988). Assessment of language disorders after right hemisphere damage. *International Journal of Language & Communication Disorders*, 23(2), 111–125. 10.3109/13682828809019881

- Burgess C, & Chiarello C (1996). Neurocognitive mechanisms underlying metaphor comprehension and other figurative language. *Metaphor and Symbolic Activity*, 11(1), 67–84. 10.1207/s15327868ms1101_4
- Caillies S, Bertot V, Motte J, Raynaud C, & Abely M (2014). Social cognition in ADHD: Irony understanding and recursive theory of mind. *Research in Developmental Disabilities*, 35(11), 3191–3198. 10.1016/j.ridd.2014.08.002 [PubMed: 25155741]
- Campbell DW, Wallace MG, Modirrousta M, Polimeni JO, McKeen NA, & Reiss JP (2015). The neural basis of humour comprehension and humour appreciation: The roles of the temporoparietal junction and superior frontal gyrus. *Neuropsychologia*, 79, 10–20. 10.1016/j.neuropsychologia.2015.10.013 [PubMed: 26474740]
- Cardillo ER, McQuire M, & Chatterjee A (2018). Selective Metaphor Impairments After Left, Not Right, Hemisphere Injury. *Frontiers in Psychology*, 9. 10.3389/fpsyg.2018.02308
- Cardillo R, Mammarella IC, Demurie E, Giofrè D, & Roeyers H (2021). Pragmatic Language in Children and Adolescents With Autism Spectrum Disorder: Do Theory of Mind and Executive Functions Have a Mediating Role? *Autism Research*, 14(5), 932–945. 10.1002/aur.2423 [PubMed: 33111475]
- Champagne-Lavau M, & Stip E (2010). Pragmatic and executive dysfunction in schizophrenia. *Journal of Neurolinguistics*, 23(3), 285–296. 10.1016/j.jneuroling.2009.08.009
- Chan Y-C, & Lavalée JP (2015a). Temporo-parietal and fronto-parietal lobe contributions to theory of mind and executive control: An fMRI study of verbal jokes. *Frontiers in Psychology*, 6. 10.3389/fpsyg.2015.01285
- Chan Y-C, & Lavalée JP (2015b). Temporo-parietal and fronto-parietal lobe contributions to theory of mind and executive control: An fMRI study of verbal jokes. *Frontiers in Psychology*, 6. 10.3389/fpsyg.2015.01285
- Channon S, & Watts M (2003). Pragmatic language interpretation after closed head injury: Relationship to executive functioning. *Cognitive Neuropsychiatry*, 8(4), 243–260. 10.1080/13546800344000002 [PubMed: 16571564]
- Chen E, Widick P, & Chatterjee A (2008). Functional–anatomical organization of predicate metaphor processing. *Brain and Language*, 107(3), 194–202. 10.1016/j.bandl.2008.06.007 [PubMed: 18692890]
- Chen X, Affourtit J, Ryskin R, Regev TI, Norman-Haignere S, Jouravlev O, Malik-Moraleda S, Kean H, Varley R, & Fedorenko E (2021). The human language system does not support music processing (p. 2021.06.01.446439). 10.1101/2021.06.01.446439
- Chen X, Lu B, & Yan C-G (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping*, 39(1), 300–318. 10.1002/hbm.23843 [PubMed: 29024299]
- Chow HM, Kaup B, Raabe M, & Greenlee MW (2008). Evidence of fronto-temporal interactions for strategic inference processes during language comprehension. *NeuroImage*, 40(2), 940–954. 10.1016/j.neuroimage.2007.11.044 [PubMed: 18201911]
- Citron FMM, & Goldberg AE (2014). Metaphorical Sentences Are More Emotionally Engaging than Their Literal Counterparts. *Journal of Cognitive Neuroscience*, 26(11), 2585–2595. 10.1162/jocn_a_00654 [PubMed: 24800628]
- Citron FMM, Güsten J, Michaelis N, & Goldberg AE (2016). Conventional metaphors in longer passages evoke affective brain response. *NeuroImage*, 139, 218–230. 10.1016/j.neuroimage.2016.06.020 [PubMed: 27346546]
- Colston HL, & O'Brien J (2000). Contrast and pragmatics in figurative language: Anything understatement can do, irony can do better. *Journal of Pragmatics*, 32(11), 1557–1583. 10.1016/S0378-2166(99)00110-1
- Coulson S, & Williams RF (2005). Hemispheric asymmetries and joke comprehension. *Neuropsychologia*, 43(1), 128–141. 10.1016/j.neuropsychologia.2004.03.015 [PubMed: 15488912]
- Dai RH, Chen H-C, Chan YC, Wu C-L, Li P, Cho SL, & Hu J-F (2017). To Resolve or Not To Resolve, that Is the Question: The Dual-Path Model of Incongruity Resolution and Absurd Verbal Humor by fMRI. *Frontiers in Psychology*, 8. 10.3389/fpsyg.2017.00498

- Deen B, & Freiwald WA (2021). Parallel systems for social and spatial reasoning within the cortical apex (p. 2021.09.23.461550). 10.1101/2021.09.23.461550
- 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. 10.1093/cercor/bhv111 [PubMed: 26048954]
- Delis DC, Wapner W, Gardner H, & Moses JA (1983). The Contribution of the Right Hemisphere to the Organization of Paragraphs. *Cortex*, 19(1), 43–50. 10.1016/S0010-9452(83)80049-5 [PubMed: 6851590]
- Demorest A, Silberstein L, Gardner H, & Winner E (1983). Telling it as it isn't: Children's understanding of figurative language. *British Journal of Developmental Psychology*, 1(2), 121–134. 10.1111/j.2044-835X.1983.tb00550.x
- Devlin J, Chang M-W, Lee K, & Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs]. <http://arxiv.org/abs/1810.04805>
- 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. *The Journal of Neuroscience*, 40(23), 4536–4550. 10.1523/JNEUROSCI.2036-19.2020 [PubMed: 32317387]
- Diaz MT, & Hogstrom LJ (2011a). The Influence of Context on Hemispheric Recruitment during Metaphor Processing. *Journal of Cognitive Neuroscience*, 23(11), 3586–3597. 10.1162/jocn_a_00053 [PubMed: 21568642]
- Diaz MT, & Hogstrom LJ (2011b). The Influence of Context on Hemispheric Recruitment during Metaphor Processing. *Journal of Cognitive Neuroscience*, 23(11), 3586–3597. 10.1162/jocn_a_00053 [PubMed: 21568642]
- DiNicola LM, Braga RM, & Buckner RL (2020). Parallel distributed networks dissociate episodic and social functions within the individual. *Journal of Neurophysiology*, 123(3), 1144–1179. 10.1152/jn.00529.2019 [PubMed: 32049593]
- DiNicola LM, & Buckner RL (2021). Precision estimates of parallel distributed association networks: Evidence for domain specialization and implications for evolution and development. *Current Opinion in Behavioral Sciences*, 40, 120–129. 10.1016/j.cobeha.2021.03.029 [PubMed: 34263017]
- 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. 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. 10.1016/j.neuron.2013.09.015 [PubMed: 24094101]
- Duncan J, & Owen AM (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10), 475–483. 10.1016/S0166-2236(00)01633-7 [PubMed: 11006464]
- Dworetzky A, Seitzman BA, Adeyemo B, Neta M, Coalson RS, Petersen SE, & Gratton C (2021). Probabilistic mapping of human functional brain networks identifies regions of high group consensus. *NeuroImage*, 237, 118164. 10.1016/j.neuroimage.2021.118164 [PubMed: 34000397]
- Egorova N, Shtyrov Y, & Pulvermüller F (2016). Brain basis of communicative actions in language. *NeuroImage*, 125, 857–867. 10.1016/j.neuroimage.2015.10.055 [PubMed: 26505303]
- Eickhoff SB, Bzdok D, Laird AR, Kurth F, & Fox PT (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, 59(3), 2349–2361. 10.1016/j.neuroimage.2011.09.017 [PubMed: 21963913]
- Eickhoff SB, Laird AR, Grefkes C, Wang LE, Zilles K, & Fox PT (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, 30(9), 2907–2926. 10.1002/hbm.20718 [PubMed: 19172646]
- Eklund A, Nichols TE, & Knutsson H (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900–7905. 10.1073/pnas.1602413113

- Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, & Peters TM (1993). 3D statistical neuroanatomical models from 305 MRI volumes. 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, 1813–1817 vol.3. 10.1109/NSSMIC.1993.373602
- Eviatar Z, & Just MA (2006). Brain correlates of discourse processing: An fMRI investigation of irony and conventional metaphor comprehension. *Neuropsychologia*, 44(12), 2348–2359. 10.1016/j.neuropsychologia.2006.05.007 [PubMed: 16806316]
- Fairchild S, & Papafragou A (2021). The Role of Executive Function and Theory of Mind in Pragmatic Computations. *Cognitive Science*, 45(2), e12938. 10.1111/cogs.12938 [PubMed: 33616218]
- Farkas AH, Trotti RL, Edge EA, Huang L-Y, Kasowski A, Thomas OF, Chlan E, Granros MP, Patel KK, & Sabatinelli D (2021). Humor and emotion: Quantitative meta analyses of functional neuroimaging studies. *Cortex*, 139, 60–72. 10.1016/j.cortex.2021.02.023 [PubMed: 33836303]
- Fedorenko E (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, 5. 10.3389/fpsyg.2014.00335
- 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. 10.1016/j.cobeha.2021.02.023
- Fedorenko E, Behr MK, & 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. 10.1073/pnas.1112937108
- Fedorenko E, & Blank IA (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, 24(4), 270–284. 10.1016/j.tics.2020.01.001 [PubMed: 32160565]
- Fedorenko E, Blank IA, Siegelman M, & Mineroff Z (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203, 104348. 10.1016/j.cognition.2020.104348 [PubMed: 32569894]
- Fedorenko E, Duncan J, & Kanwisher N (2012). Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. *Current Biology*, 22(21), 2059–2062. 10.1016/j.cub.2012.09.011 [PubMed: 23063434]
- 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. 10.1073/pnas.1315235110
- 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. 10.1152/jn.00032.2010 [PubMed: 20410363]
- Fedorenko E, Scott TL, Brunner P, Coon WG, 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. 10.1073/pnas.1612132113
- 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. 10.1177/09637214211046955 [PubMed: 35295820]
- Fedorenko E, & Thompson-Schill SL (2014). Reworking the language network. *Trends in Cognitive Sciences*, 18(3), 120–126. 10.1016/j.tics.2013.12.00 [PubMed: 24440115]
- 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.
- 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. 10.1016/j.bandl.2017.04.005 [PubMed: 28527316]
- Ferstl EC, Neumann J, Bogler C, & von Cramon DY (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, 29(5), 581–593. 10.1002/hbm.20422 [PubMed: 17557297]
- Ferstl EC, & von Cramon DY (2001). The role of coherence and cohesion in text comprehension: An event-related fMRI study. *Cognitive Brain Research*, 11(3), 325–340. 10.1016/S0926-6410(01)00007-6 [PubMed: 11339984]

- Ferstl EC, Walther K, Guthke T, & von Cramon DY (2005). Assessment of Story Comprehension Deficits After Brain Damage. *Journal of Clinical and Experimental Neuropsychology*, 27(3), 367–384. 10.1080/13803390490515784 [PubMed: 15969358]
- Filik R, urcan A, Ralph-Nearman C, & Pitiot A (2019). What is the difference between irony and sarcasm? An fMRI study. *Cortex*, 115, 112–122. 10.1016/j.cortex.2019.01.025 [PubMed: 30807881]
- Fischl B, Rajendran N, Busa E, Augustinack J, Hinds O, Yeo BTT, Mohlberg H, Amunts K, & Zilles K (2008). Cortical Folding Patterns and Predicting Cytoarchitecture. *Cerebral Cortex*, 18(8), 1973–1980. 10.1093/cercor/bhm225 [PubMed: 18079129]
- Friese U, Rutschmann R, Raabe M, & Schmalhofer F (2008). Neural Indicators of Inference Processes in Text Comprehension: An Event-related Functional Magnetic Resonance Imaging Study. *Journal of Cognitive Neuroscience*, 20(11), 2110–2124. 10.1162/jocn.2008.20141 [PubMed: 18416672]
- Frost MA, & Goebel R (2012). Measuring structural–functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. *NeuroImage*, 59(2), 1369–1381. 10.1016/j.neuroimage.2011.08.035 [PubMed: 21875671]
- Gagnon L, Goulet P, Giroux F, & Joanne Y (2003). Processing of metaphoric and non-metaphoric alternative meanings of words after right- and left-hemispheric lesion. *Brain and Language*, 87(2), 217–226. 10.1016/S0093-934X(03)00057-9 [PubMed: 14585291]
- Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, & Frith CD (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. 10.1016/S0028-3932(99)00053-6 [PubMed: 10617288]
- Genovese CR, Lazar NA, & Nichols T (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15(4), 870–878. 10.1006/nimg.2001.1037 [PubMed: 11906227]
- Gernsbacher MA, & Robertson RRW (1999). The role of suppression in figurative language comprehension. *Journal of Pragmatics*, 31(12), 1619–1630. 10.1016/S0378-2166(99)00007-7 [PubMed: 25520540]
- Geschwind N (1970). The Organization of Language and the Brain. *Science*. 10.1126/science.170.3961.940
- Gibbs RW Jr. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, 34(4), 457–486. 10.1016/S0378-2166(01)00046-7
- Giora R, Zaidel E, Soroker N, Batori G, & Kasher A (2000). Differential effect of right- and left-hemisphere damage on understanding sarcasm and metaphor. *Metaphor and Symbol*, 15(1–2), 63–83. 10.1207/S15327868MS151&2_5
- Glucksberg S, & McGlone MS (2001). *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press, USA.
- Goel V, & Dolan RJ (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition*, 93(3), B109–B121. 10.1016/j.cognition.2004.03.001 [PubMed: 15178381]
- Gordon EM, & Nelson SM (2021). Three types of individual variation in brain networks revealed by single-subject functional connectivity analyses. *Current Opinion in Behavioral Sciences*, 40, 79–86. 10.1016/j.cobeha.2021.02.014
- Graesser AC, Singer M, & Trabasso T (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371–395. 10.1037/0033-295X.101.3.371 [PubMed: 7938337]
- Gratton C, & Braga RM (2021). Editorial overview: Deep imaging of the individual brain: past, practice, and promise. *Current Opinion in Behavioral Sciences*, 40, iii–vi. 10.1016/j.cobeha.2021.06.011
- Grice HP (1975). Logic and conversation. *Syntax and Semantics*, 3, 41–58.
- Happé FGE (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2), 101–119. 10.1016/0010-0277(93)90026-R [PubMed: 8243028]
- Happé F, Brownell H, & Winner E (1999). Acquired ‘theory of mind’ impairments following stroke. *Cognition*, 70(3), 211–240. 10.1016/S0010-0277(99)00005-0 [PubMed: 10384736]

- Heidlmayr K, Weber K, Takashima A, & Hagoort P (2020). No title, no theme: The joined neural space between speakers and listeners during production and comprehension of multi-sentence discourse. *Cortex*, 130, 111–126. 10.1016/j.cortex.2020.04.035 [PubMed: 32652339]
- Hellbernd N, & Sammler D (2018). Neural bases of social communicative intentions in speech. *Social Cognitive and Affective Neuroscience*, 13(6), 604–615. 10.1093/scan/nsy034 [PubMed: 29771359]
- Herold R, Varga E, Hajnal A, Hamvas E, Berecz H, Tóth B, & Tényi T (2018). Altered Neural Activity during Irony Comprehension in Unaffected First-Degree Relatives of Schizophrenia Patients—An fMRI Study. *Frontiers in Psychology*, 8. 10.3389/fpsyg.2017.02309
- Hillert DG, & Bura GT. (2009). The neural substrates of spoken idiom comprehension. *Language and Cognitive Processes*, 24(9), 1370–1391. 10.1080/01690960903057006
- Hu J, Levy R, & Zaslavsky N (2021). Scalable pragmatic communication via self-supervision. *Proceedings of the 2021 ICML Workshop on Self-Supervised Learning for Reasoning and Perception*. arXiv:2108.05799.
- Hu J, Floyd S, Jouravlev O, Fedorenko E, & Gibson E (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. arXiv. 10.48550/arXiv.2212.06801
- Hugdahl K, Raichle ME, Mitra A, & Specht K (2015). On the existence of a generalized non-specific task-dependent network. *Frontiers in Human Neuroscience*, 9. 10.3389/fnhum.2015.00430
- Ianni GR, Cardillo ER, McQuire M, & Chatterjee A (2014). Flying under the radar: Figurative language impairments in focal lesion patients. *Frontiers in Human Neuroscience*, 8. 10.3389/fnhum.2014.00871
- Ivanova AA, Srikant S, Sueoka Y, Kean HH, Dhamala R, O'Reilly U-M, Bers MU, & Fedorenko E (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *ELife*, 9, e58906. 10.7554/eLife.58906 [PubMed: 33319744]
- Jackendoff R (2002). Semantic and conceptual foundations. In *Foundations of language: Brain, meaning, grammar, evolution* (pp. 267–293). Oxford University 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. 10.1016/j.neuroimage.2015.11.025 [PubMed: 26589334]
- Jang G, Yoon S, Lee S-E, Park H, Kim J, Ko JH, & Park H-J (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage*, 81, 61–72. 10.1016/j.neuroimage.2013.05.027 [PubMed: 23684863]
- Joanette Y, Goulet P, Hannequin D, & Boeglin J (1990). Right hemisphere and verbal communication. Springer-Verlag Publishing. 10.1007/978-1-4612-4460-8
- Julian JB, 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. 10.1016/j.neuroimage.2012.02.055 [PubMed: 22398396]
- Jung-Beeman M (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, 9(11), 512–518. 10.1016/j.tics.2005.09.009 [PubMed: 16214387]
- Klepousniotou E, & Baum SR (2005). Unilateral brain damage effects on processing homonymous and polysemous words. *Brain and Language*, 93(3), 308–326. 10.1016/j.bandl.2004.10.011 [PubMed: 15862856]
- Klooster N, McQuire M, Grossman M, McMillan C, Chatterjee A, & Cardillo E (2020). The Neural Basis of Metaphor Comprehension: Evidence from Left Hemisphere Degeneration. *Neurobiology of Language*, 1(4), 474–491. 10.1162/nol_a_00022 [PubMed: 37215584]
- Kristensen LB, Wang L, Petersson KM, & Hagoort P (2013). The Interface Between Language and Attention: Prosodic Focus Marking Recruits a General Attention Network in Spoken Language Comprehension. *Cerebral Cortex*, 23(8), 1836–1848. 10.1093/cercor/bhs164 [PubMed: 22763170]
- Kuperberg GR, Lakshmanan BM, Caplan DN, & Holcomb PJ (2006). Making sense of discourse: An fMRI study of causal inferencing across sentences. *NeuroImage*, 33(1), 343–361. 10.1016/j.neuroimage.2006.06.001 [PubMed: 16876436]
- Kuperberg GR, McGuire PK, Bullmore ET, Brammer MJ, Rabe-Hesketh S, Wright IC, Lythgoe DJ, Williams SCR, & David AS (2000). Common and Distinct Neural Substrates for Pragmatic,

- Semantic, and Syntactic Processing of Spoken Sentences: An fMRI Study. *Journal of Cognitive Neuroscience*, 12(2), 321–341. 10.1162/089892900562138 [PubMed: 10771415]
- Lauro LJR, Tettamanti M, Cappa SF, & Papagno C (2008). Idiom Comprehension: A Prefrontal Task? *Cerebral Cortex*, 18(1), 162–170. 10.1093/cercor/bhm042 [PubMed: 17490991]
- Lee SS, & Dapretto M (2006). Metaphorical vs. literal word meanings: FMRI evidence against a selective role of the right hemisphere. *NeuroImage*, 29(2), 536–544. 10.1016/j.neuroimage.2005.08.003 [PubMed: 16165371]
- Licea-Haquet GL, Velásquez-Upegui EP, Holtgraves T, & Giordano M (2019). Speech act recognition in Spanish speakers. *Journal of Pragmatics*, 141, 44–56. 10.1016/j.pragma.2018.12.013
- Lin N, Yang X, Li J, Wang S, Hua H, Ma Y, & Li X (2018). Neural correlates of three cognitive processes involved in theory of mind and discourse comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, 18(2), 273–283. 10.3758/s13415-018-0568-6
- Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, Kean H, Jouravlev O, Rakocevic L, Pritchett B, Siegelman M, Hoeflin C, Pongos A, Blank IA, Struhl MK, Ivanova A, Shannon S, Sathe A, Hoffmann M, Nieto-Castañón A, & Fedorenko E (2022). Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, 9(1), Article 1. 10.1038/s41597-022-01645-3
- Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, Nieto-Castañón A, and Fedorenko E (in preparation). A probabilistic atlas for the Multiple Demand (MD) network based on data from 691 individuals performing a spatial working memory localizer task.
- Lisofsky N, Kazzer P, Heekeren HR, & Prehn K (2014). Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. *Neuropsychologia*, 61, 113–122. 10.1016/j.neuropsychologia.2014.06.001 [PubMed: 24929201]
- Mahowald K, & Fedorenko E (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*, 139, 74–93. 10.1016/j.neuroimage.2016.05.073 [PubMed: 27261158]
- Martín-Loeches M, Casado P, Hernández-Tamames JA, & Álvarez-Linera J (2008). Brain activation in discourse comprehension: A 3t fMRI study. *NeuroImage*, 41(2), 614–622. 10.1016/j.neuroimage.2008.02.047 [PubMed: 18394923]
- Mashal N, Faust M, & Hendler T (2005a). The role of the right hemisphere in processing nonsalient metaphorical meanings: Application of Principal Components Analysis to fMRI data. *Neuropsychologia*, 43(14), 2084–2100. 10.1016/j.neuropsychologia.2005.03.019 [PubMed: 16243053]
- Mashal N, Faust M, & Hendler T (2005b). The role of the right hemisphere in processing nonsalient metaphorical meanings: Application of Principal Components Analysis to fMRI data. *Neuropsychologia*, 43(14), 2084–2100. 10.1016/j.neuropsychologia.2005.03.019 [PubMed: 16243053]
- Mashal N, Faust M, Hendler T, & Jung-Beeman M (2007). An fMRI investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain and Language*, 100(2), 115–126. 10.1016/j.bandl.2005.10.005 [PubMed: 16290261]
- Mason RA, & Just MA (2011). Differentiable cortical networks for inferences concerning people's intentions versus physical causality. *Human Brain Mapping*, 32(2), 313–329. 10.1002/hbm.21021 [PubMed: 21229617]
- Mather M, Cacioppo JT, & Kanwisher N (2013). How fMRI Can Inform Cognitive Theories. *Perspectives on Psychological Science*, 8(1), 108–113. 10.1177/1745691612469037 [PubMed: 23544033]
- Matsui T, Nakamura T, Utsumi A, Sasaki AT, Koike T, Yoshida Y, Harada T, Tanabe HC, & Sadato N (2016). The role of prosody and context in sarcasm comprehension: Behavioral and fMRI evidence. *Neuropsychologia*, 87, 74–84. 10.1016/j.neuropsychologia.2016.04.031 [PubMed: 27157883]
- Matthews D, Biney H, & Abbot-Smith K (2018). Individual Differences in Children's Pragmatic Ability: A Review of Associations with Formal Language, Social Cognition, and Executive Functions. *Language Learning and Development*, 14(3), 186–223. 10.1080/15475441.2018.1455584

- McDonald S, & Pearce S (1998). Requests That Overcome Listener Reluctance: Impairment Associated with Executive Dysfunction in Brain Injury. *Brain and Language*, 61(1), 88–104. 10.1006/brln.1997.1846 [PubMed: 9448933]
- Moher D, Liberati A, Tetzlaff J, Altman DG, & Group TP (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, 6(7), e1000097. 10.1371/journal.pmed.1000097 [PubMed: 19621072]
- Monti MM, Parsons LM, & Osherson DN (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences*, 106(30), 12554–12559. 10.1073/pnas.0902422106
- Monti MM, Parsons LM, & Osherson DN (2012). Thought beyond language: Neural dissociation of algebra and natural language. *Psychological Science*, 23(8), 914–922. 10.1177/0956797612437427 [PubMed: 22760883]
- Myers PS, Linebaugh CW, 1981. Comprehension of idiomatic expressions by right-hemisphere-damaged adults. In Brookshire RH (Ed.), *Clinical aphasiology: Conference proceedings* pp. 254–261. Minneapolis: BRK.
- Myers P (1998). *Right hemisphere damage: Disorders of communication and cognition*. San Diego, CA: Singular.
- Nagels A, Kauschke C, Schrauf J, Whitney C, Straube B, & Kircher Ti. (2013). Neural substrates of figurative language during natural speech perception: An fMRI study. *Frontiers in Behavioral Neuroscience*, 7. 10.3389/fnbeh.2013.00121
- Nieto-Castañón A, & Fedorenko E (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*, 63(3), 1646–1669. 10.1016/j.neuroimage.2012.06.065 [PubMed: 22784644]
- Obert A, Gierski F, Calmus A, Flucher A, Portefaix C, Pierot L, Kaladjian A, & Caillies S (2016). Neural Correlates of Contrast and Humor: Processing Common Features of Verbal Irony. *PLOS ONE*, 11(11), e0166704. 10.1371/journal.pone.0166704 [PubMed: 27851821]
- Obert A, Gierski F, Calmus A, Portefaix C, Declercq C, Pierot L, & Caillies S (2014). Differential bilateral involvement of the parietal gyrus during predicative metaphor processing: An auditory fMRI study. *Brain and Language*, 137, 112–119. 10.1016/j.bandl.2014.08.002 [PubMed: 25193417]
- Oliveri M, Romero L, & Papagno C (2004). Left But Not Right Temporal Involvement in Opaque Idiom Comprehension: A Repetitive Transcranial Magnetic Stimulation Study. *Journal of Cognitive Neuroscience*, 16(5), 848–855. 10.1162/089892904970717 [PubMed: 15200712]
- Ouden D-B den, Dickey MW., Anderson C& Christianson K. (2016). Neural correlates of early-closure garden-path processing: Effects of prosody and plausibility. *The Quarterly Journal of Experimental Psychology*, 69(5), 926–949. 10.1080/17470218.2015.1028416 [PubMed: 25801097]
- Papagno C, & Genoni A (2004). The role of syntactic competence in idiom comprehension: A study on aphasic patients. *Journal of Neurolinguistics*, 17(5), 371–382. 10.1016/j.jneuroling.2003.11.002
- Paunov AM (2018). *FMRI studies of the relationship between language and theory of mind in adult cognition* [Thesis, Massachusetts Institute of Technology]. <https://dspace.mit.edu/handle/1721.1/121828>
- Paunov AM, Blank IA, & 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. 10.1152/jn.00619.2018 [PubMed: 30601693]
- Paunov AM, Blank IA, Jouravlev O, Mineroff Z, Gallée J, & Fedorenko E (2021). Differential tracking of linguistic vs. Mental state content in naturalistic stimuli by language and Theory of Mind (ToM) brain networks (p. 2021.04.28.441724). 10.1101/2021.04.28.441724
- Perrone-Bertolotti M, Dohen M, Lævenbruck H, Sato M, Pichat C, & Baciu M (2013). Neural correlates of the perception of contrastive prosodic focus in French: A functional magnetic resonance imaging study. *Human Brain Mapping*, 34(10), 2574–2591. 10.1002/hbm.22090 [PubMed: 22488985]

- Piñango MM, Zhang M, Foster-Hanson E, Negishi M, Lacadie C, & Constable RT (2017). Metonymy as Referential Dependency: Psycholinguistic and Neurolinguistic Arguments for a Unified Linguistic Treatment. *Cognitive Science*, 41(S2), 351–378. 10.1111/cogs.12341 [PubMed: 26887916]
- Poldrack RA (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2), 59–63. 10.1016/j.tics.2005.12.004 [PubMed: 16406760]
- Poldrack RA (2011). Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. *Neuron*, 72(5), 692–697. 10.1016/j.neuron.2011.11.001 [PubMed: 22153367]
- Pomp J, Bestgen A-K, Schulze P, Müller CJ, Citron FMM, Suchan B, & Kuchinke L (2018). Lexical olfaction recruits olfactory orbitofrontal cortex in metaphorical and literal contexts. *Brain and Language*, 179, 11–21. 10.1016/j.bandl.2018.02.001 [PubMed: 29482170]
- Premack D, & Woodruff G (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. 10.1017/S0140525X00076512
- Rapp AM (2019). Comprehension of Metaphors and Idioms: An Updated Meta-analysis of Functional Magnetic Resonance Imaging Studies. In de Zubicaray GI & Schiller NO (Eds.), *The Oxford Handbook of Neurolinguistics* (pp. 710–735). Oxford University Press. 10.1093/oxfordhb/9780190672027.013.28
- Rapp AM, Erb M, Grodd W, Bartels M, & Markert K (2011). Neural correlates of metonymy resolution. *Brain and Language*, 119(3), 196–205. 10.1016/j.bandl.2011.07.004 [PubMed: 21889196]
- Rapp AM, Leube DT, Erb M, Grodd W, & Kircher TTJ (2004). Neural correlates of metaphor processing. *Cognitive Brain Research*, 20(3), 395–402. 10.1016/j.cogbrainres.2004.03.017 [PubMed: 15268917]
- Rapp AM, Mutschler DE, & Erb M (2012). Where in the brain is nonliteral language? A coordinate-based meta-analysis of functional magnetic resonance imaging studies. *NeuroImage*, 63(1), 600–610. 10.1016/j.neuroimage.2012.06.022 [PubMed: 22759997]
- Rapp AM, Mutschler DE, Wild B, Erb M, Lengsfeld I, Saur R, & Grodd W (2010). Neural correlates of irony comprehension: The role of schizotypal personality traits. *Brain and Language*, 113(1), 1–12. 10.1016/j.bandl.2009.11.007 [PubMed: 20071019]
- Regev TI, Affourtit J, Chen X, Schipper AE, Bergen L, Mahowald K, & Fedorenko E (2021). High-level language brain regions are sensitive to sub-lexical regularities (p. 2021.06.11.447786). 10.1101/2021.06.11.447786
- Reyes-Aguilar A, Valles-Capetillo E, & Giordano M (2018). A Quantitative Meta-analysis of Neuroimaging Studies of Pragmatic Language Comprehension: In Search of a Universal Neural Substrate. *Neuroscience*, 395, 60–88. 10.1016/j.neuroscience.2018.10.043 [PubMed: 30414881]
- Rints A, McAuley T, & Nilsen ES (2015). Social Communication Is Predicted by Inhibitory Ability and ADHD Traits in Preschool-Aged Children: A Mediation Model. *Journal of Attention Disorders*, 19(10), 901–911. 10.1177/1087054714558873 [PubMed: 25477018]
- Samur D, Lai VT, Hagoort P, & Willems RM (2015a). Emotional context modulates embodied metaphor comprehension. *Neuropsychologia*, 78, 108–114. 10.1016/j.neuropsychologia.2015.10.003 [PubMed: 26449989]
- Samur D, Lai VT, Hagoort P, & Willems RM (2015b). Emotional context modulates embodied metaphor comprehension. *Neuropsychologia*, 78, 108–114. 10.1016/j.neuropsychologia.2015.10.003 [PubMed: 26449989]
- 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. 10.1016/S1053-8119(03)00230-1 [PubMed: 12948738]
- Saxe R, Moran JM, Scholz J, & Gabrieli J (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, 1(3), 229–234. 10.1093/scan/nsl034 [PubMed: 18985110]
- Saxe R, & Powell LJ (2006). It’s the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, 17(8), 692–699. 10.1111/j.1467-9280.2006.01768.x [PubMed: 16913952]

- Schmidt GL, & Seger CA (2009). Neural correlates of metaphor processing: The roles of figurativeness, familiarity and difficulty. *Brain and Cognition*, 71(3), 375–386. 10.1016/j.bandc.2009.06.001 [PubMed: 19586700]
- Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown EN, & Saxe R (2009). Distinct Regions of Right Temporo-Parietal Junction Are Selective for Theory of Mind and Exogenous Attention. *PLOS ONE*, 4(3), e4869. 10.1371/journal.pone.0004869 [PubMed: 19290043]
- Schuil KDI, Smits M, & Zwaan RA (2013). Sentential Context Modulates the Involvement of the Motor Cortex in Action Language Processing: An fMRI Study. *Frontiers in Human Neuroscience*, 7. 10.3389/fnhum.2013.00100
- Scott TL, 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. 10.1080/17588928.2016.1201466 [PubMed: 27386919]
- Searle J (1979). *Metaphor*. In Ortony A (Ed.), *Metaphor and thought*. New York: Cambridge University Press.
- Shain C, Blank IA, van Schijndel M, Schuler W, & Fedorenko E (2020). FMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138, 107307. 10.1016/j.neuropsychologia.2019.107307 [PubMed: 31874149]
- Shain C, Paunov A, Chen X, Lipkin B, & Fedorenko E (2022). No evidence of theory of mind reasoning in the human language network. *bioRxiv*. 10.1101/2022.07.18.500516
- Shashidhara S, Mitchell DJ, Erez Y, & Duncan J (2019). Progressive Recruitment of the Frontoparietal Multiple-demand System with Increased Task Complexity, Time Pressure, and Reward. *Journal of Cognitive Neuroscience*, 31(11), 1617–1630. 10.1162/jocn_a_01440 [PubMed: 31274390]
- Shibata M, Abe J, Itoh H, Shimada K, & Umeda S (2011). Neural processing associated with comprehension of an indirect reply during a scenario reading task. *Neuropsychologia*, 49(13), 3542–3550. 10.1016/j.neuropsychologia.2011.09.006 [PubMed: 21930137]
- Shibata M, Abe J, Terao A, & Miyamoto T (2007). Neural mechanisms involved in the comprehension of metaphoric and literal sentences: An fMRI study. *Brain Research*, 1166, 92–102. 10.1016/j.brainres.2007.06.040 [PubMed: 17662699]
- Shibata M, Terasawa Y, & Umeda S (2014). Integration of cognitive and affective networks in humor comprehension. *Neuropsychologia*, 65, 137–145. 10.1016/j.neuropsychologia.2014.10.025 [PubMed: 25447374]
- Shibata M, Toyomura A, Itoh H, & Abe J (2010). Neural substrates of irony comprehension: A functional MRI study. *Brain Research*, 1308, 114–123. 10.1016/j.brainres.2009.10.030 [PubMed: 19853585]
- Shibata M, Toyomura A, Motoyama H, Itoh H, Kawabata Y, & Abe J (2012). Does simile comprehension differ from metaphor comprehension? A functional MRI study. *Brain and Language*, 121(3), 254–260. 10.1016/j.bandl.2012.03.006 [PubMed: 22534570]
- Siegal M, Carrington J, & Radel M (1996). Theory of Mind and Pragmatic Understanding Following Right Hemisphere Damage. *Brain and Language*, 53(1), 40–50. 10.1006/brln.1996.0035 [PubMed: 8722898]
- Smith DM, Perez DC, Porter A, Dworetzky A, & Gratton C (2021). Light through the fog: Using precision fMRI data to disentangle the neural substrates of cognitive control. *Current Opinion in Behavioral Sciences*, 40, 19–26. 10.1016/j.cobeha.2020.12.004 [PubMed: 33553511]
- Smith N, & Levy R (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).
- Smith V, Duncan J, & Mitchell DJ (2021). Roles of the Default Mode and Multiple-Demand Networks in Naturalistic versus Symbolic Decisions. *The Journal of Neuroscience*, 41(10), 2214–2228. 10.1523/JNEUROSCI.1888-20.2020 [PubMed: 33472829]
- Sperber D, & Wilson D (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.
- Spotorno N, Koun E, Prado J, Van Der Henst J-B, & Noveck IA (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1), 25–39. 10.1016/j.neuroimage.2012.06.046 [PubMed: 22766167]

- Stemmer B, Giroux F, & Joanne Y (1994). Production and Evaluation of Requests by Right Hemisphere Brain-Damaged Individuals. *Brain and Language*, 47(1), 1–31. 10.1006/brln.1994.1040 [PubMed: 7922473]
- Stringaris AK, Medford NC, Giampietro V, Brammer MJ, & David AS (2007). Deriving meaning: Distinct neural mechanisms for metaphoric, literal, and non-meaningful sentences. *Brain and Language*, 100(2), 150–162. 10.1016/j.bandl.2005.08.001 [PubMed: 16165201]
- Tahmasebi AM, Davis MH, Wild CJ, Rodd JM, Hakyemez H, Abolmaesumi P, & Johnsrude IS (2012). Is the Link between Anatomical Structure and Function Equally Strong at All Cognitive Levels of Processing? *Cerebral Cortex*, 22(7), 1593–1603. 10.1093/cercor/bhr205 [PubMed: 21893681]
- Thirion B, Thual A, & Pinho AL (2021). From deep brain phenotyping to functional atlas. *Current Opinion in Behavioral Sciences*, 40, 201–212. 10.1016/j.cobeha.2021.05.004
- Tian F, Hou Y, Zhu W, Dietrich A, Zhang Q, Yang W, Chen Q, Sun J, Jiang Q, & Cao G (2017). Getting the Joke: Insight during Humor Comprehension – Evidence from an fMRI Study. *Frontiers in Psychology*, 8. 10.3389/fpsyg.2017.01835
- Tompkins CA (1990). Knowledge and Strategies for Processing Lexical Metaphor after Right or Left Hemisphere Brain Damage. *Journal of Speech, Language, and Hearing Research*, 33(2), 307–316. 10.1044/jshr.3302.307
- Turkeltaub PE, Eden GF, Jones KM, & Zeffiro TA (2002). Meta-Analysis of the Functional Neuroanatomy of Single-Word Reading: Method and Validation. *NeuroImage*, 16(3, Part A), 765–780. 10.1006/nimg.2002.1131 [PubMed: 12169260]
- Turkeltaub PE, Eickhoff SB, Laird AR, Fox M, Wiener M, & Fox P (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, 33(1), 1–13. 10.1002/hbm.21186 [PubMed: 21305667]
- van Ackeren MJ, 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. 10.1162/jocn_a_00274 [PubMed: 22849399]
- van Ackeren MJ, Smaragdi A, & Rueschemeyer S-A (2016). Neuronal interactions between mentalising and action systems during indirect request processing. *Social Cognitive and Affective Neuroscience*, 11(9), 1402–1410. 10.1093/scan/nsw062 [PubMed: 27131039]
- Van Lancker DR, & Kempler D (1987). Comprehension of familiar phrases by left-but not by right-hemisphere damaged patients. *Brain and Language*, 32(2), 265–277. 10.1016/0093-934X(87)90128-3 [PubMed: 2446699]
- Varga E, Schnell Zs., Tényi T., Németh N., Simon M, Hajnal A., Horváth RA, Hamvas E, Járαι., Fekete, & Herold R. (2014). Compensatory effect of general cognitive skills on non-literal language processing in schizophrenia: A preliminary study. *Journal of Neurolinguistics*, 29, 1–16. 10.1016/j.jneuroling.2014.01.001
- Varga E, Simon M, Tényi T, Schnell Zs., Hajnal A, Orsi G., Dóczi T, Komoly S, Janszky J, Füredi R, Hamvas E, Fekete S, & Herold R. (2013). Irony comprehension and context processing in schizophrenia during remission – A functional MRI study. *Brain and Language*, 126(3), 231–242. 10.1016/j.bandl.2013.05.017 [PubMed: 23867921]
- Vartanian O (2012). Dissociable neural systems for analogy and metaphor: Implications for the neuroscience of creativity. *British Journal of Psychology*, 103(3), 302–316. 10.1111/j.2044-8295.2011.02073.x [PubMed: 22804698]
- Virtue S, Haberman J, Clancy Z, Parrish T, & Jung Beeman M (2006). Neural activity of inferences during story comprehension. *Brain Research*, 1084(1), 104–114. 10.1016/j.brainres.2006.02.053 [PubMed: 16574079]
- Virtue S, Parrish T, & Jung-Beeman M (2008). Inferences during Story Comprehension: Cortical Recruitment Affected by Predictability of Events and Working Memory Capacity. *Journal of Cognitive Neuroscience*, 20(12), 2274–2284. 10.1162/jocn.2008.20160 [PubMed: 18457505]
- Vrticka P, Black JM, & Reiss AL (2013). The neural basis of humour processing. *Nature Reviews Neuroscience*, 14(12), 860–868. 10.1038/nrn3566 [PubMed: 24169937]
- Wakusawa K, Sugiura M, Sassa Y, Jeong H, Horie K, Sato S, Yokoyama H, Tsuchiya S, Inuma K, & Kawashima R (2007). Comprehension of implicit meanings in social situations involving irony:

- A functional MRI study. *NeuroImage*, 37(4), 1417–1426. 10.1016/j.neuroimage.2007.06.013 [PubMed: 17689103]
- Wang AT, Lee SS, Sigman M, & Dapretto M (2006). Developmental changes in the neural basis of interpreting communicative intent. *Social Cognitive and Affective Neuroscience*, 1(2), 107–121. 10.1093/scan/nsi018 [PubMed: 18985123]
- Wehbe L, Blank IA, Shain C, Futrell R, Levy R, von der Malsburg T, Smith N, Gibson E, & Fedorenko E (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, 31(9), 4006–4023. 10.1093/cercor/bhab065 [PubMed: 33895807]
- Weylman ST, Brownell HH, Roman M, & Gardner H (1989). Appreciation of indirect requests by left- and right-brain-damaged patients: The effects of verbal context and conventionality of wording. *Brain and Language*, 36(4), 580–591. 10.1016/0093-934X(89)90087-4 [PubMed: 2470462]
- Whitney P, Ritchie BG, & Crane RS (1992). The effect of foregrounding on readers' use of predictive inferences. *Memory & Cognition*, 20(4), 424–432. 10.3758/BF03210926 [PubMed: 1495404]
- Whyte EM, & Nelson KE (2015). Trajectories of pragmatic and nonliteral language development in children with autism spectrum disorders. *Journal of Communication Disorders*, 54, 2–14. 10.1016/j.jcomdis.2015.01.001 [PubMed: 25638464]
- Willems RM, de Boer M, de Ruiter JP, Noordzij ML, Hagoort P, & Toni I (2010). A Dissociation Between Linguistic and Communicative Abilities in the Human Brain. *Psychological Science*, 21(1), 8–14. 10.1177/0956797609355563 [PubMed: 20424015]
- 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. 10.1016/0010-0277(83)90004-5 [PubMed: 6681741]
- Winner E, Brownell H, Happé F, Blum A, & Pincus D (1998). Distinguishing Lies from Jokes: Theory of Mind Deficits and Discourse Interpretation in Right Hemisphere Brain-Damaged Patients. *Brain and Language*, 62(1), 89–106. 10.1006/brln.1997.1889 [PubMed: 9570881]
- Winner E, & Gardner H (1977). The comprehension of metaphor in brain-damaged patients. *Brain*, 100(4), 717–729. 10.1093/brain/100.4.717 [PubMed: 608117]
- Wittgenstein L (1953). *Philosophical investigations*. Oxford: Basil Blackwell.
- Woolgar A, Parr A, Cusack R, Thompson R, Nimmo-Smith I, Torralva T, Roca M, Antoun N, Manes F, & Duncan J (2010). Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 107(33), 14899–14902. 10.1073/pnas.1007928107
- Wu X, Guo T, Zhang C, Hong T-Y, Cheng C-M, Wei P, Hsieh J-C, & Luo J (2021). From “Aha!” to “Haha!” Using Humor to Cope with Negative Stimuli. *Cerebral Cortex*, 31(4), 2238–2250. 10.1093/cercor/bhaa357 [PubMed: 33258955]
- Yang FG, Edens J, Simpson C, & Krawczyk DC (2009). Differences in task demands influence the hemispheric lateralization and neural correlates of metaphor. *Brain and Language*, 111(2), 114–124. 10.1016/j.bandl.2009.08.006 [PubMed: 19781756]
- Yang J (2014). The role of the right hemisphere in metaphor comprehension: A meta-analysis of functional magnetic resonance imaging studies. *Human Brain Mapping*, 35(1), 107–122. 10.1002/hbm.22160 [PubMed: 22936560]
- Yang J, Li P, Fang X, Shu H, Liu Y, & Chen L (2016). Hemispheric involvement in the processing of Chinese idioms: An fMRI study. *Neuropsychologia*, 87, 12–24. 10.1016/j.neuropsychologia.2016.04.029 [PubMed: 27143223]
- Yang J, & Shu H (2016). Involvement of the Motor System in Comprehension of Non-Literal Action Language: A Meta-Analysis Study. *Brain Topography*, 29(1), 94–107. 10.1007/s10548-015-0427-5 [PubMed: 25681159]
- Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, & Wager TD (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8), 665–670. 10.1038/nmeth.1635 [PubMed: 21706013]
- Yarkoni T, Speer NK, & Zacks JM (2008). Neural substrates of narrative comprehension and memory. *NeuroImage*, 41(4), 1408–1425. 10.1016/j.neuroimage.2008.03.062 [PubMed: 18499478]

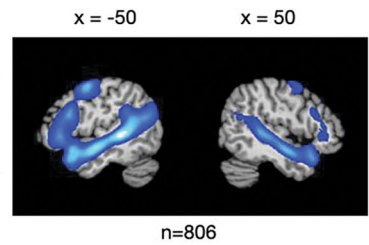
- Yi YG, Kim DY, Shim WH, Oh JY, Kim SH, & Kim HS (2017). Neural correlates of Korean proverb processing: A functional magnetic resonance imaging study. *Brain and Behavior*, 7(10), e00829. 10.1002/brb3.829 [PubMed: 29075575]
- Zaidel E, Kasher A, Soroker N, & Batori G (2002). Effects of Right and Left Hemisphere Damage on Performance of the “Right Hemisphere Communication Battery.” *Brain and Language*, 80(3), 510–535. 10.1006/brln.2001.2612 [PubMed: 11896655]
- Zempleni M-Z, Haverkort M, Renken R, & Stowe A, L. (2007). Evidence for bilateral involvement in idiom comprehension: An fMRI study. *NeuroImage*, 34(3), 1280–1291. 10.1016/j.neuroimage.2006.09.049 [PubMed: 17141528]

Language

Sentences

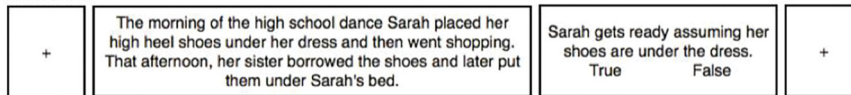


Non-words

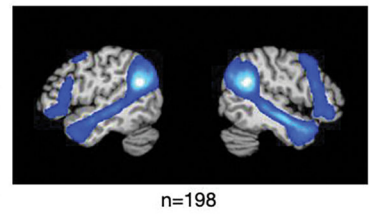
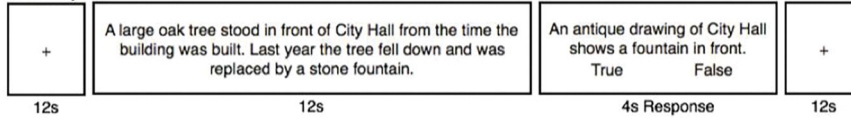


Theory of Mind

False Belief

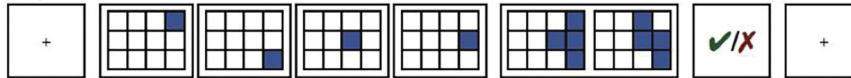


False Physical



Multiple Demand

Easy



Hard

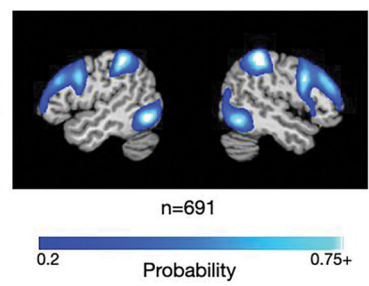
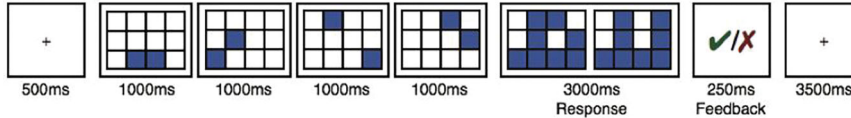


Figure 1.

The three functional localizer paradigms (language, ToM, MD) and the resulting probabilistic functional atlases. The maps illustrate, for each voxel, the proportion of participants for whom that voxel belongs to the top 10% of localizer-responsive voxels. Despite the apparent “overlap” between language and ToM regions at the level of the probabilistic atlases, these networks show little to no overlap in individual participants (see Blank et al., 2014; Paunov et al., 2018; Braga et al., 2020).

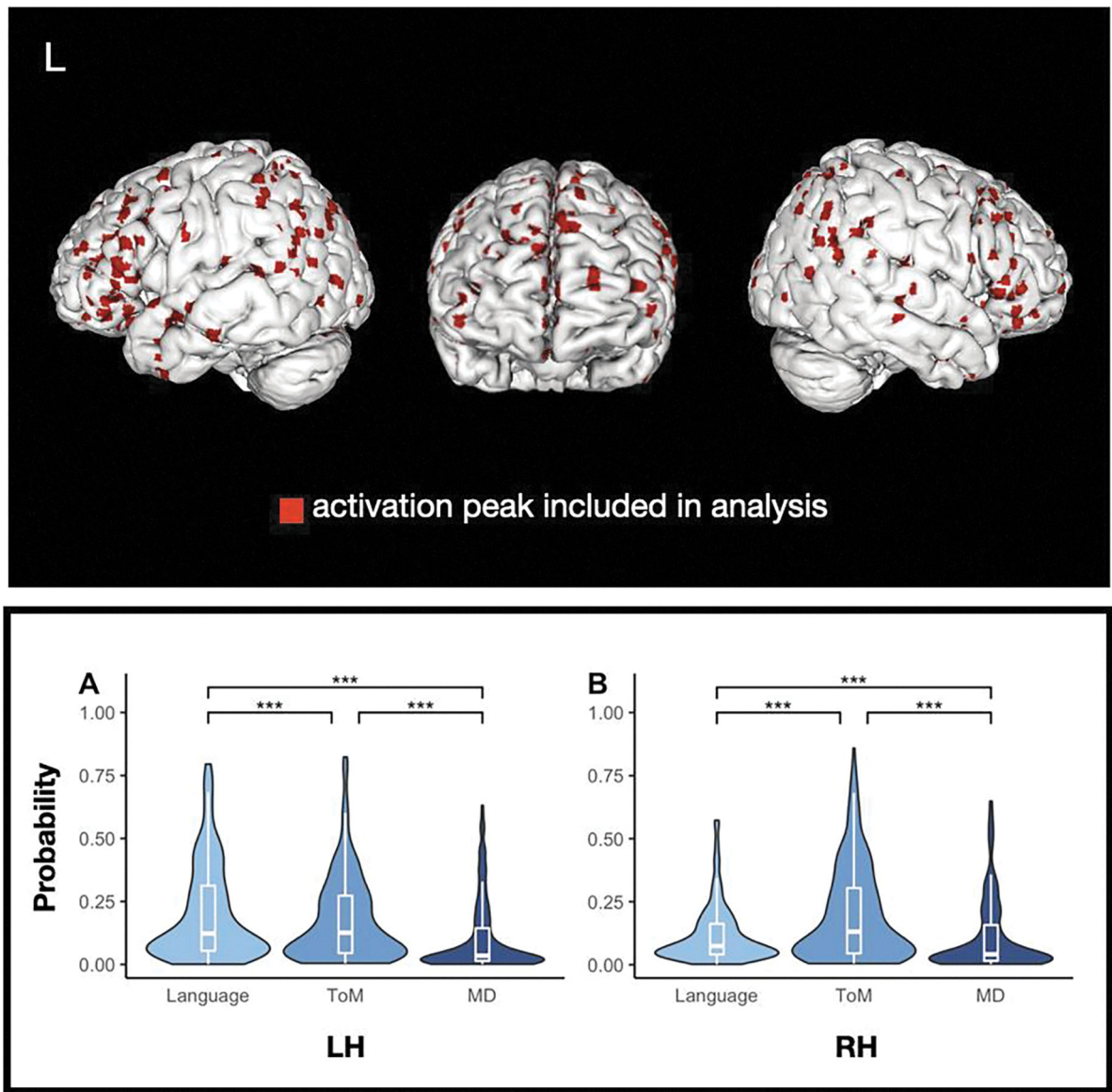


Figure 2. Individual peaks. The top panel displays individual activation peaks plotted on the smoothed (width = 2; kernel size = 3) cortical surface of a high-resolution structural MRI scan in MNI space. The bottom panel displays the distribution of nonzero network probabilities associated with all individual peaks, by hemisphere. Significance of median differences between networks was assessed via a permutation test; all p s < .0001.

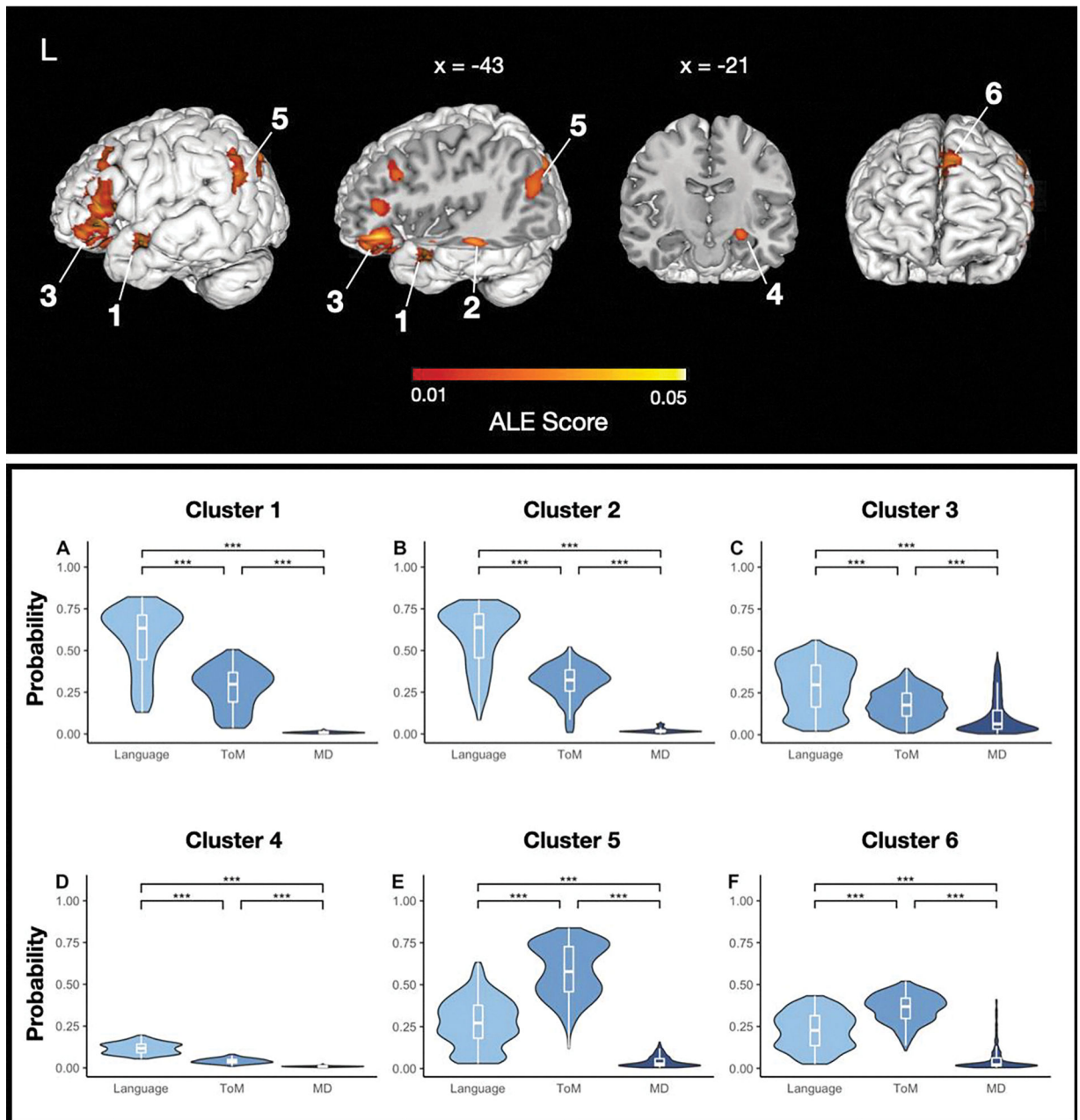


Figure 3.

Activation likelihood estimation (ALE) results. The top panel displays ALE scores associated with each of the 6 significant ALE clusters, plotted on the smoothed (width = 2; kernel size = 3) cortical surface of a high-resolution structural MRI scan in MNI space (cluster-forming threshold = $p < 0.001$ uncorrected, cluster-level family-wise error (FWE) = $p < 0.05$, 1000 permutations). The bottom panel displays the distribution of nonzero network probabilities associated with all voxels contained within each cluster. Significance of median differences between networks was assessed via a permutation test; all $ps < .0001$.

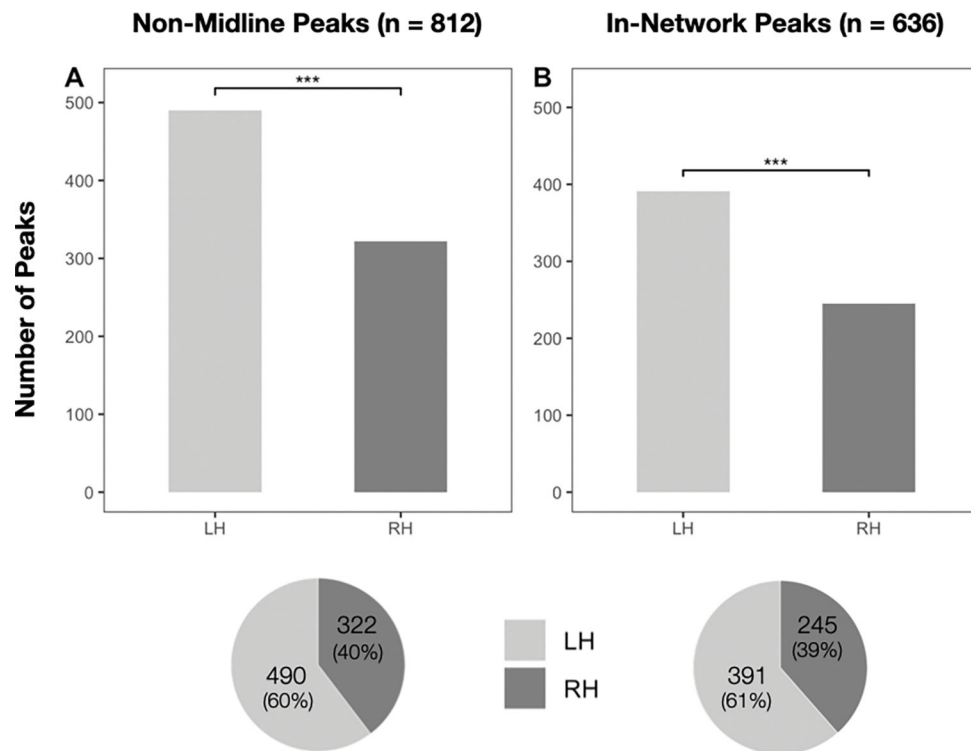


Figure 4. Peaks by hemisphere. In the top panel, A displays the number of non-midline peaks in the left vs. right hemisphere. B displays the left vs. right breakdown of non-midline peaks with an network probability of at least 0.10 in at least one network. Significance was assessed via logistic mixed effects regression; all p s < .001. Pie charts displaying the proportion of activation peaks in the left vs. right hemisphere are shown in the bottom panel.

Table 1.

Distribution of studies and peaks across linguistic phenomena. (One of the 74 studies included separate contrasts for non-sarcastic irony and sarcasm and is therefore counted twice in the table.) Studies on prosody that focused on the role of emotionally charged prosodic cues were excluded.

Linguistic phenomenon	Number of studies	Number of peaks
Humor	8	121
Idiom	6	54
Indirect Speech	10	160
Irony	10	68
Metaphor	19	201
Metonymy	2	10
Prosody	4	64
Proverb	1	9
Sarcasm	2	10
Text Coherence	13	128
TOTAL	74 unique studies	825

Table 2.

Details on the design and timing of the language, ToM and MD localizer tasks.

	Language	ToM	MD
Design	Blocked	Long-event-related	Blocked
Length of trial	4.8–18 s	14 s	8–9 s
Trials per block	1–5	N/A	2–4
Blocks/events per run	12–18	10	10–12
Blocks/events per condition per run	4–16	5	5–6
Length of run	336–504 s	272 s	288–448 s
Runs	2–8	1–2	2–4
Versions of the design	10	1	3

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3.

Network probabilities for the six significant ALE clusters. The number of voxels in each cluster is listed under “Size (in voxels)”. The “Broad Anatomical Area” column lists the macroanatomical region that overlaps the most with each cluster (as determined by the ALE). In the “Language,” “ToM,” and “MD” columns, the first value represents the network probability associated with that cluster’s center peak and the second value represents the median network probability of all voxels contained in the cluster. In bold, we highlighted for each cluster the network that has the highest network probabilities.

Cluster	Size (in voxels)	Hemisphere	Broad Anatomical Area	x	y	z	Language	ToM	MD
1	283	LH	STG	-54	-1	-16	0.743 / 0.634	0.348 / 0.298	0.001 / 0.009
2	357	LH	MTG	-55	-33	-3	0.738 / 0.638	0.394 / 0.323	0.016 / 0.016
3	1,695	LH	IFG	-47	26	3	0.424 / 0.298	0.273 / 0.177	0.030 / 0.064
4	176	LH	Amygdala	-23	-10	-17	0.136 / 0.117	0.030 / 0.040	0.004 / 0.009
5	528	LH	STG	-49	-63	29	0.197 / 0.272	0.611 / 0.578	0.016 / 0.028
6	454	LH	MedFG	-3	51	31	0.355 / 0.227	0.449 / 0.369	0.012 / 0.022