Brain & Language 171 (2017) 52–61

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Effects of contextual relevance on pragmatic inference during conversation: An fMRI study

Wangshu Feng [a,b], Yue Wu [a,b], Catherine Jan [a,b], Hongbo Yu [a,b], Xiaoming Jiang [a,b,c], Xiaolin Zhou [a,b,d,e,f,*]

[a] Center for Brain and Cognitive Sciences, Peking University, Beijing 100871, China
[b] School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China
[c] School of Communication Sciences and Disorders, McGill University, Montréal H3A 1G1, Canada
[d] Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing 100871, China
[e] Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China
[f] PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 16 November 2016
Revised 24 March 2017
Accepted 24 April 2017

Keywords:
Pragmatic inference
Indirect reply
Conversational implicature
Contextual relevance
fMRI
Functional connectivity

ABSTRACT

Contextual relevance, which is vital for understanding conversational implicatures (CI), engages both the frontal-temporal language and theory-of-mind networks. Here we investigate how contextual relevance affects CI processing and regulates the connectivity between CI-processing-related brain regions. Participants listened to dialogues in which the level of contextual relevance to dialogue-final utterance (reply) was manipulated. This utterance was either direct, indirect but relevant, irrelevant with contextual hint, or irrelevant with no contextual hint. Results indicated that compared with direct replies, indirect replies showed increased activations in bilateral IFG, bilateral MTG, bilateral TPJ, dmPFC, and precuneus, and increased connectivity between rTPJ/dmPFC and both IFG and MTG. Moreover, irrelevant replies activated right MTG along an anterior-posterior gradient as a function of the level of irrelevance. Our study provides novel evidence concerning how the language and theory-of-mind networks interact for pragmatic inference and how the processing of CI is modulated by level of contextual relevance.

1. Introduction

Imagine that a student asks her friend “Is Dr. Smith in the office now”; her friend might say “His car is parked outside the building”. In this case, her friend did not directly answer the student’s question but successfully conveyed her speculation that Dr. Smith was in the office at that time. In daily conversations, such dissociation of the non-literal meaning implicated by the speaker (conversational implicature or CI) and the literal meaning decoded directly from linguistic expressions is quite common. The hearer often has to infer speaker’s intention beyond literal meaning from context-mediated information. Such inferential processing, or pragmatic inference, starts from the results of initial semantic analysis, uses knowledge about context, and elicits a conclusion beyond these available hints (Grice, 1975). According to the Relevance Theory, the principle of relevance between utterances and its context drives the inferential process in understanding what the speaker intends to communicate. The context of an utterance supports to recover the utterance interpretation that crosses the relevance threshold. Thus, more processing efforts are required in order to infer speaker’s true intentions conveyed by an utterance with less contextual relevance (Carston, 2004; Sperber & Wilson, 1986).

Neuroimaging studies found that the neural correlates of pragmatic inference can be divided into two categories: the core language network and the extra-language areas (Hagoort, 2013; Hagoort & Levinson, 2014). Firstly, the core language network for semantic processing is found to be associated with understanding the implicated meanings of speakers’ utterances (pragmatic inference). Non-literal language comprehension involves language regions in both left and right hemispheres (Rapp, Leube, Erb, Grodd, & Kircher, 2004, 2007; Rapp, Mutschler, & Erb, 2012); in particular, right hemisphere is predominantly sensitive to coherence between the utterance and its context in discourse processing (Kuperberg, Lakshmanan, Caplan, & Holcomb, 2006; Menenti, Petersson, Scheeringa, & Hagoort, 2009; Nieuwland, 2012). The higher activations in language cortex (e.g. bilateral inferior frontal gyri) reflect listener’s extra efforts to fill the semantic gap between the literal meaning of linguistic expressions and its discourse contexts (Ferstl & von Cramon, 2001; Siebörger, Ferstl, & von Cramon,

* Corresponding author at: School of Psychological and Cognitive Sciences, Peking University, 5 Yiheyuan Road, Beijing 100871, China.
E-mail address: xz104@pku.edu.cn (X. Zhou).

http://dx.doi.org/10.1016/j.bandl.2017.04.005
0093-934X/

2007; Uchiyama et al., 2012). To this end, the listener has to integrate both the textual and derived information in order to construct the most optimal relevance in communication (Sperber & Wilson, 1986), which requires broadly related semantic activation, selection and integration. Secondly, extra-language areas are also found to be associated with understanding nonliteral meaning (Bohrn, Altmann, & Jacobs, 2012; Ferstl, Neumann, Bogler, & von Cramon, 2008; Rapp et al., 2012). Mounting studies indicate that the processing of pragmatic inference depends on Theory-of-Mind (ToM) related network (e.g. Bašnáková, Weber, Petersson, van Berkum, & Hagoort, 2014; Egorova, Pulvermüller, & Shtyrov, 2014; van Ackeren, Casasanto, Bekkering, Hagoort, & Ruesschemeyer, 2012). The ToM network, typically consisting of bilateral temporo-parietal junction (TPJ), medial prefrontal cortex (mPFC) and precuneus, is generally known to be involved in making inferences about mental states of others (Koster-Hale & Saxe, 2013; Mar, 2011; Premack & Woodruff, 1978).

Previous neuroimaging studies have revealed the neural substrate of contextual relevance in communication by comparing indirect speech with direct speech (Bašnáková, van Berkum, Weber, & Hagoort, 2015; Bašnáková et al., 2014; Shibata, Abe, Itoh, Shimada, & Umeda, 2011; van Ackeren, Smaragdi, & Rueschemeyer, 2016; van Ackeren et al., 2012). An fMRI study on indirect requests (van Ackeren et al., 2012) presented participants with an auditory sentence paired with a visual scene picture. The sentence-picture combinations in the critical condition could be interpreted as indirect requests, while those in the control conditions could only be interpreted as plausible statements. This study found that the cortical motor system and ToM network were activated in comprehending indirect requests, relative to the control sentences. Another set of studies investigated the neural correlates of comprehending indirect replies by using natural dialogues as experimental material, in which the critical utterance was interpreted either as an indirect or direct reply to a certain question. By comparing indirect replies to direct replies, Bašnáková et al. (2014, 2015) observed greater activations in bilateral IFG, right (and left) MTG, mPFC extending into SMA, and right (and left) TPJ; Shibata et al. (2011) found activations in bilateral IFG, right MTG, mPFC, and precuneus; and van Ackeren et al. (2016) found activations in left IFG, mPFC, right insula, and bilateral caudate.

Jang and colleagues went further to manipulate levels of contextual relevance and to investigate its subtle influence on inferring the implicated meanings (Jang, Yoon, Lee, Park, & Kim, 2013). This study used materials with explicit, moderately implicit and highly implicit answers to a simple question. For example, for a yes-no question "Is Dr. Smith in his office now?" there were three types of answers: (1) "Dr. Smith is in his office now" (explicit); (2) "Dr. Smith's car is parked outside the building" (moderately implicit); and (3) "The black car is parked outside the building" (highly implicit). Participants were asked to read dialogues and to decide whether the answer to the question meant 'yes' or 'no'. The fMRI results showed that left anterior temporal lobe, left angular gyrus, and left posterior MTG showed stronger activations in both the moderately and highly implicit conditions than in the explicit condition. Comprehension of highly implicit answers involved activations in additional regions including left IFG, left mPFC, left posterior cingulate cortex and right anterior temporal lobe. However, a problem with this study is that it did not control the literal linguistic expressions between direct vs. indirect speech conditions. The differences incurred from the comparison between the two conditions might be caused by the processing of different syntactic, semantic, or low-level pragmatic information (e.g. referents of pronouns), rather than by the generation of CI.

Built on these previous studies, the present study aimed to (1) further characterize neural substrates underlying the processing of indirect replies with various contextual relevance by using more strictly controlled language materials, and (2) reveal how the connectivity between brain regions involved in pragmatic inferential processing is modulated by the contextual relevance. To this end, we measured the listener's neural responses during listening to replies with various levels of relevance in natural spoken dialogues. In each dialogue, depending on the cover story and the preceding question, which represented situational context and immediate context respectively (see Table 1), a certain final utterance could serve as a direct reply (DR), a relevant reply (RR), an irrelevant reply with contextual hint (IRC) or an irrelevant reply with no contextual hint (IRNC). The cover stories of each set in the DR, RR and IRC conditions were the same, but different preceding questions were created to satisfy the reply in the different conditions. For example, the utterance "Nowadays, people are really beginning to enjoy opera" served as a direct reply to the question "Do you think that more people are beginning to like opera", as an indirect but relevant reply to the question "Do you think that the audience enjoyed my opera performance", or as a literally irrelevant reply to the question "Do you think the audience will vote for my performance". There was an overlap in the word usage in the

Table 1
Examples of the four different types of replies, translated into English. The critical utterances are in bold, and main differences among these four conditions are italics.

Condition		Cover Story	Dialogue
Direct reply (DR)		After completing her opera performance at a music competition, *the opera singer* comes out of her dressing room and sees her friend. The following is the dialogue between the singer and her friend[a]	Q: Do you think that more people are *beginning to like opera*? **A: Nowadays, people are really beginning to enjoy opera**
Indirect Reply	Relevant reply (RR)	After completing her opera performance at a music competition, *the opera singer* comes out of her dressing room and sees her friend. The following is the dialogue between the singer and her friend	Q: Do you think that the audience *liked my opera performance*? **A: Nowadays, people are really beginning to enjoy opera.**
	Irrelevant reply, with contextual hint (IRC)	After completing her opera performance at a music competition, *the opera singer* comes out of her dressing room and sees her friend. The following is the dialogue between the singer and her friend.	Q: Do you think the audience will *vote for my performance*? **A: Nowadays, people are really beginning to enjoy opera**
	Irrelevant reply, without contextual hint (IRNC)	After completing musical performance at a music competition, *the singer* comes out of her dressing room and sees her friend. The following is the dialogue between the singer and her friend	Q: Do you think the audience will *vote for my performance*? **A: Nowadays, people are really beginning to enjoy opera**

[a] Please note that it was implicit to the Chinese participants that the audience's approval was directly related to the musical competition, as the audience voted on the musical performance (e.g., American Idol).

reply (critical utterance) and the preceding question in the RR condition, whereas there was no overlap in the word usage in the replies and the preceding questions in the IRC/IRNC condition. Thus, understanding the implicatures of the irrelevant replies (in the IRC/IRNC condition) needs more semantic and inferential processing to connect the utterance with the preceding question, compared with relevant replies (in the RR condition). The IRC and IRNC conditions shared the same dialogue, but unlike the cover story in the IRNC condition, that in the IRC condition provided information that linked the reply with its immediate context. For example, the cover story of the IRC condition, "After completing her opera performance at a music competition, the opera singer comes out of her dressing room and sees her friend. The following is the dialogue between the singer and her friend", contained critical information that the dialogue protagonist was an opera singer; this was to construct relevance between "vote for performance" in the question and "enjoy opera" in the reply. But the cover story of the IRNC condition did not contain this critical information. Thus, under the IRNC condition, participants were more likely to rely on out-of-context knowledge to infer the speaker's intention, relative to under the IRC condition. Within each set of items in our materials (i.e., one set is composed of four items), all indirect replies communicated the same attitude or intention (social motive), while different sets of stimulus materials involved different types of social motives, such as to save one's face, to provide more information, or to display modesty. For this "opera" example, the speaker would use an indirect reply if he/she is not absolutely sure about the personal tastes of the audience (the RR condition) or the results of the competition (the IRC and IRNC condition), while, at the same time, still wanting to convey positive evaluation about the singer's performance and provide more information in support of his/her attitude.

Common to all the experimental conditions (DR, RR, IRC, and IRNC) is the fact that the preceding question sets up a strong expectation for a yes/no answer and the final utterance gives a definite answer (for our example, the critical utterance is equivalent to a 'yes' to its preceding question in all the four scenarios). Given that inferring the speaker's meaning intentionally is a necessary condition for the generation of CI (Bach, 2006), we used a listening comprehension task in which participants were asked to listen to each dialogue and to make binary judgment as to what the latter speaker really wanted to say with his/her utterance.

Based on previous studies, we predicted that the reaction time (RT) in binary judgment would significantly increase over the DR, RR, IRC and IRNC conditions.

We assume that at least two additional cognitive processes would occur during comprehending speakers' implicated meaning. First, listeners need to integrate all accessible information, both textual and derived, to fill the semantic gap and to establish linguistic coherence between the critical utterance and its context (Jang et al., 2013). This process requires broadly related semantic activation, selection and integration. Since we manipulated the semantic relationship between the reply and its context, the imaging results would show that the frontal-temporal language network is activated in comprehending indirect replies and may be differentially involved in the indirect replies with varying contextual relevance. Second, understanding the social motives behind the indirect replies depends on the inferential process in relation to ToM (Bašnáková et al., 2014, 2015; van Ackeren et al., 2012). Hence our imaging results would show that the ToM related brain network would be engaged in understanding indirect reply.

Rather than simply replicating previous studies in another language (Chinese), the current study went further to investigate how the brain regions work together to give rise to pragmatic inference. Using the psychophysiological interaction (PPI) analysis, Spotorno and his colleagues found that compared to the literal meaning, understanding ironic meaning involved enhanced effective functional connectivity between mPFC (as a representative region of the ToM network) and bilateral IFG (Spotorno, Koun, Prado, Van Der Henst, & Noveck, 2012). Using dynamic causal modelling, van Ackeren et al. (2016) showed that mPFC received input from left IFG during understanding indirect speech. For the current study, we used PPI analysis to extend the connectivity map between the language network and ToM network during understanding indirect speech.

2. Method

2.1. Participants

Results of twenty-three university students (12 females, mean age 22.4 years, SD = 1.97) were analyzed. Four additional participants were excluded from data analysis because of poor task performance (three standard deviations lower than average in response accuracy) or excessive head movement (>3 mm). All participants were right-handed native Chinese speakers with normal or corrected-to-normal vision and had no history of neurological or psychiatric disorders. None of them suffered from any hearing or language disorder. They all provided written informed consent. The study was approved by the Ethics Committee of the Department of Psychology at Peking University.

2.2. Design and materials

One hundred and sixty sets of scenarios were selected by a pretest (see below). There were four scenarios in each set that shared the same critical utterances and had different situational or immediate contexts (see Table 1). Each scenario was composed of three components: a cover story that briefly introduced communication circumstance, a yes-no question, and a direct/indirect reply to that question as the critical utterance. Each scenario corresponded to one particular experimental condition. There were four different conditions used in the experimental design: a direct reply condition (DR) as a baseline and three types of indirect reply conditions. The direct reply requires minimal amount of inference to understand the speaker's intention. The three indirect reply conditions were created and identified according to the literally contextual relevance between the critical utterance and its context: the relevant reply (RR), the irrelevant reply with contextual hint (IRC) and the irrelevant reply without contextual hint (IRNC). Specifically, the critical utterance in the RR condition was relevant literally to the question in the content, but the critical utterance in the IRC/IRNC condition (these two conditions shared the same dialogue) seemed irrelevant to the preceding question in the content. The only difference between the IRC and IRNC conditions was in the cover story, which gave hints to the relevance between the question and its reply in the IRC condition, but not in the IRNC condition. Within each set of items, all indirect replies communicate the same social motive (in the forms of either attitude or intention), while different sets of stimuli may involve different types of social motives. For the selected materials (160 sets in total), indirect replies in 41 sets were used for face-saving situations, involving polite refusals not to offend the person asking the question; 64 sets were used to provide more information that supported his/her attitude/argument than just a simple "yes/no" (as in the above "opera" example); 28 sets were used to convey high praise (e.g. "Do you like my poetry?", "I can already recite your poetry"); and 27 sets were used to protect one's own face (e.g. "How did you do on the mid-term?", "The mid-term was full of Olympic math questions") or to display modesty (e.g., "Did you finish preparing for the test this afternoon?", "This test does not cover much content."). Note that, all these social motives outlined above are very natural in the Chinese culture. Each reply, either direct or indirect, gave a definite answer. Half of the replies answered "yes" to the question while half of them answered "no".

To confirm the manipulation of contextual relevance, the cosine similarity between the context (cover story + question) and the reply was calculated using a Latent Semantic Analysis (LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) trained on a comprehensive corpus (the full text of Chinese Wikipedia; https://dumps.wikimedia.org). A repeated-measure analysis of variance (ANOVA) over the mean cosine similarity for the four conditions showed a significant main effect of condition, $F(3,477) = 46.11$, $p < 0.001$ (see Fig. 1A), with decreasing semantic similarity between the context and the reply over the DR (mean = 0.34, SD = 0.28), the RR (mean = 0.25, SD = 0.25), the IRC (mean = 0.21, SD = 0.22), and the IRNC (mean = 0.17, SD = 0.18) conditions. Differences between the four conditions were all significant ($ps < 0.01$, with Bonferroni correction for multiple comparisons). These results objectively quantified the differences between the four conditions.

All scenarios were presented in auditory form to achieve more natural simulation of conversation environment. Six female and six male Chinese native speakers took part in recording audio of the cover stories and dialogues. The cover story, the question and the answer were recorded by different speakers for a particular scenario. All dialogues occurred between a male speaker and a female speaker. All audio materials were digitized at 16-bit/11.0 kHz sampling rate and equated for the maximum sound intensity.

2.3. Pretests

We administered a pretest to access the level of indirectness of the dialogues and to select the final set of stimuli. Thirty-four participants who did not participate in the fMRI scanning took part in this pretest. We generated 172 sets of scenarios and divided them into four lists according to a Latin-square procedure. Each list was rated by 8/9 participants. During this pretest, participants were asked to listen carefully to each scenario, make binary judgment as to what the speaker really wanted to say with his/her utterance, and then rate how directly the reply answered the question on a 7-point visual analog scale (1 representing the most indirect reply and 7 representing the most direct reply).

To minimize the influence of ambiguity during comprehending indirect reply, we excluded 12 sets of scenarios because two participants or more disagreed with others on the speaker's meaning of the critical utterance in one or more scenarios. Thus, one hundred and sixty sets of scenarios were selected for the formal experiment. Consistent with the above LSA, a repeated-measure ANOVA over the mean rating scores for the four conditions showed a significant main effect of experimental condition, $F(3,99) = 25.83$, $p < 0.001$, with decreased directness for replies in the DR (mean = 5.97, SD = 1.19), the RR (mean = 4.17, SD = 0.76), the IRC (mean = 3.96, SD = 0.77) and the IRNC (mean = 3.87, SD = 0.78) conditions. Except for the difference between the IRC and IRNC conditions, the differences between the DR, RR, IRC, and IRNC conditions were all significant ($ps < 0.05$, with Bonferroni correction for multiple comparisons). These results indicated that the dialogues with less contextual relevance were considered more indirect.

Twenty-eight participants who did not participate in either indirectness rating or fMRI scanning took part in a pilot behavioral experiment. The experimental procedure was the same as that in the fMRI experiment (see the Method Section 2.4). They were asked to make yes/no judgment as soon as possible after the presentation of the auditory stimuli and the visual cue of response ("yes" and "no" words on the screen). A repeated-measure ANOVA over the mean RTs for the four conditions showed a significant main effect of condition, $F(3,81) = 10.72$, $p < 0.001$ (see Fig. 1B), with increased RTs for replies over the DR (mean = 459 ms, SD = 144), the RR (mean = 489 ms, SD = 136), the IRC (mean = 509 ms, SD = 153), and the IRNC (mean = 547 ms, SD = 162) conditions. Except for the difference between the RR and IRC conditions, the differences between conditions were all significant ($ps < 0.05$, with Bonferroni correction for multiple comparisons).

2.4. Procedure

The fMRI scanning was divided into two sessions, lasting approximately 20 min per session. We divided all scenarios into four experimental lists according to a Latin-square procedure, with each list split further for the two sessions. Each participant received one scenario in each set. Scenarios in each list were pseudo-randomized with the restriction that no more than 3 scenarios of the same condition were presented consecutively and no more than 4 consecutive scenarios required the same response in the binary forced-choice task.

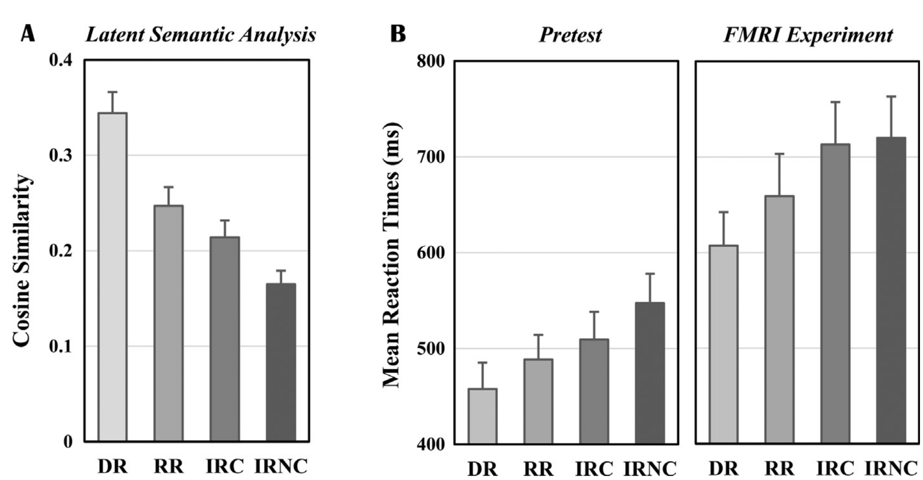

Fig. 1. (A) Semantic similarity between the context and the reply decreased as a function of experimental condition, with DR having the greatest semantic similarity and IRNC having the lowest semantic similarity in Latent Semantic Analysis (LSA). (B) Behavioral results from the test phase of the pretest (left panel) and the fMRI experiment (right panel). Mean RTs (ms) increased as a function of experimental condition (DR, RR, IRC, and IRNC). Error bars denote between-subject standard errors.

For each trial, a fixation cross was first presented at the center of the screen for a jittered duration from 1.5 to 5.5 s, followed by a blank interval for 0.1 s. Then the cover story, question and its reply was played consecutively, during which only a fixation cross was presented in the screen. There was an interval of 1 s between the cover story and the question, while there was a jittered interval from 0.5 to 1.5 s between the question and its reply. The reply was presented, immediately followed by a "yes" word on the left side of the screen and a "no" word on the right side of the screen for 3 s. Participants were asked to make binary judgments as to whether the speaker wanted to say "yes" or "no" to the preceding question. Participants were asked to carry out the judgment as accurately and as quickly as possible by pressing the appropriate button with their index or middle finger of their right hand.

Before scanning, all participants received a written instruction about the experimental procedure and completed a practice session. To investigate participants' understanding of each scenario, all fMRI participants were asked to return to the lab a few weeks later after scanning and rate the indirectness of the replies in the same way as the pretest.

2.5. Data acquisition

Functional images were acquired on a GE-MR750 3T system, using a T2*-weighted echo planar imaging (EPI) sequence, with 2000 ms repetition time, 30 ms echo time, and 90° flip angle. Each image consisted of 35 axial slices covering the whole brain. Slice thickness was 4 mm and inter-slice gap was 0.75 mm, with a 192 mm field of view (FOV), 64×64 matrix, and $3.0 \times 3.0 \times 4$ mm^3 voxel size. Head motion was minimized using pillows and cushions around the head and forehead strap.

2.6. Data analysis

For each experimental condition, task accuracy was acquired and RT was measured as the time cost to make judgment after the reply was presented. We conducted behavioral data analyses based on repeated measured ANOVA. The threshold for statistical significance was defined as $\alpha = 0.05$. P-values were adjusted with the Greenhouse-Geisser correction for nonsphericity when necessary and with Bonferroni correction for multiple comparisons.

The fMRI data were pre-processed with Statistical Parametric Mapping (SPM) software SPM8 (Welcome Department of Imaging Neuroscience, London, http://www.fil.ion.ucl.ac.uk). The first five volumes of each session were discarded to allow for stabilization of magnetization. The remaining images were time sliced and realigned to the sixth volume of the first session for head movement. A temporal high-pass filter with a cutoff frequency of 1/128 Hz was used to remove low-frequency drifts in an fMRI time series, and the mean functional image for each participant was coregistered to the EPI template provided by SPM8. Images were anatomically normalized to the MNI space (resampled to $3 \times 3 \times 3$ mm^3 isotropic voxel) by matching gray matter (Ashbumer & Friston, 2005), and smoothed with a Gaussian kernel of 6 mm full-width half-maximum (FWHM). Participants whose head movements did not exceed 3 mm were included in the final data analysis.

Statistical analysis was based on the general linear model (GLM) first at the participant level and then at the group level. Events in each trial were modeled as a boxcar function convolved with a canonical hemodynamic response function (HRF). Two models were built to collect convergent evidence. For the factorial model, we defined eight regressors in the GLM: four for the reply presentation, one for the cover story presentation, one for the question presentation, one for the judgment, and one for the button press. The reply presentation regressors were defined as the conditions of interest (i.e., DR, RR, IRC, and IRNC). For the parametric model, the four reply presentation regressors in the factorial model were combined into a single regressor, accompanied by a parametric regressor containing the participants' post-scanning rating of indirectness for each sentence. Both models additionally included six rigid body parameters to correct for the head motion artifact. The onset and duration of the regressors of interest was set to the onset and duration of each auditory utterance.

To pinpoint regions significantly activated for the conditions of interest, we first calculated the simple main effect in each condition. The participant level individual images of four conditions of interest were then fed to a flexible factorial repeated measures analysis of variance in the group level design matrix. A cortical mask that excluded the cerebellum was used at the group level, with further analyses also being conducted within this masking. We defined six contrasts: three contrasts of three types of indirect reply effect, which compared the RR, IRC and IRNC conditions with the DR condition respectively, a main contrast of indirectness, which compared pooled indirect reply conditions with the direct reply condition, and two contrasts that compared the IRC vs. RR condition, and the IRNC vs. IRC condition respectively.

A Monte Carlo simulation (Ledberg, Åkerman, & Roland, 1998) implemented in the AFNI program AlphaSim (http://afni.nimh.nih.gov/pub/dist/doc/manual/AlphaSim.pdf) was used to determine the significance criterion. Areas of activation were identified as significant only if they passed the threshold of voxel-wise $p < 0.001$ uncorrected combined with cluster level threshold of $p < 0.05$ family wise corrected, which required 17 contiguous voxels (459 mm^3) or more (unless otherwise stated).

To reveal the differential contributions of right middle temporal gyrus (rMTG; see the *Results* Section 3.2) between the three types of indirect replies, we extracted the parameter estimates from three 4 mm-radius spherical regions distributed along anterior rMTG-to-posterior rMTG (MNI coordinates of the center: [54, −4, −35], [60, −37, −8], and [66, −52, 10]). Then we computed for each region the effects of three indirect reply types (i.e. RR – DR, IRC – DR, and IRNC – DR). A three by three repeated measures ANOVA was performed to formally test whether the activation pattern in these rMTG positions was modulated by the type of indirect reply.

2.6.1. Conjunction analysis

To explore areas showing pragmatic inference effects common for different types of indirect replies, we also conducted an SPM 'conjunction null' analysis (Nichols, Brett, Andersson, Wager, & Poline, 2005) using the following contrasts: (RR-DR) ∩ (IRC-DR) ∩ (IRNC-DR) (Friston, Holmes, Price, Büchel, & Worsley, 1999). Analyses were carried out with voxel-level threshold of $p < 0.001$ uncorrected and a cluster-level threshold of $p < 0.05$, AlphaSim corrected for multiple comparisons.

2.6.2. Psychophysiological interaction analysis

Psychophysiological interaction (PPI) analysis was used to investigate the functional connectivity between two brain regions that was varied with the experimental manipulation (Friston et al., 1997). Our interest lied in functional connectivity between ToM network and semantic network that was modulated by reply indirectness. To this end, we computed a PPI map with the main contrast of indirectness and used typical ToM regions (right TPJ and dmPFC) identified in our univariate analysis as seed regions. These analyses were modeled with three regressors for PPI and six regressors for the head motion artifact. The first two regressors were the time-series data extracted from a 4 mm-radius sphere centered at the peak coordinates in the region of interest (physiological regressor), and the experimental condition (direct reply vs. indirect reply) vector convolved with a canonical HRF (psychological regressor), respectively. The third regressor was the interaction term that was computed by the physiological regressor and the psychological regressor. Analyses were carried out with the same statistical threshold as stated above.

3. Results

3.1. Behavioral results

In the online task, participants correctly responded to 98.2% of all trials in the DR condition, 97.2% in the RR condition, 97.2% in the IRC condition, and 96.1% in the IRNC condition on average. Repeated-measures ANOVA did not show a significant main effect of condition, $F(3,66) = 1.50$, $p = 0.228$. Trials with incorrect response were excluded from the following behavioral data and fMRI data analyses.

Repeated-measures ANOVA revealed a significant main effect for RTs, $F(3,66) = 7.59$, $p = 0.001$ (see Fig. 1B), with the RT increasing in the order of the DR (mean = 607 ms, SD = 170 ms), RR (mean = 659 ms, SD = 212 ms), IRC (mean = 713 ms, SD = 213 ms), and IRNC (mean = 720 ms, SD = 208 ms) conditions. Except the differences between DR and RR and the differences between IRC and IRNC, the differences between conditions were all significant, *ps* < 0.05, with Bonferroni correction for multiple comparisons. The slower responses to the last two kinds of irrelevant reply suggested that comprehension of these dialogues involves more complex pragmatic inferential processing. The RTs in the pilot study and fMRI scanning showed slightly different patterns in that the difference between the IRC and IRNC conditions was reduced in fMRI scanning. This could be due to the increased RTs in fMRI scanning which could have created a ceiling effect.

For the off-line rating, a repeated-measure ANOVA over the mean rating scores for the four conditions showed a significant main effect of condition, $F(3,66) = 171.37$, $p < 0.001$, with decreased directness for replies in the DR (mean = 6.40, SD = 0.28), the RR (mean = 3.91, SD = 0.54), the IRC (mean = 3.66, SD = 0.52) and the IRNC (mean = 3.72, SD = 0.57) conditions. Except for the difference between the IRC and IRNC, the differences between the DR, RR, IRC, and IRNC conditions were all significant (*ps* < 0.05, with Bonferroni correction for multiple comparisons). These results of indirectness rating repeated the pattern in the pretest.

3.2. fMRI results

A comparison between pooled indirect replies and direct replies (DR, as the baseline condition) identified activations in bilateral temporo-parietal junction (TPJ; BA 37/39), precuneus (BA 17/23), bilateral middle temporal gyri (MTG; BA 20/21), and bilateral inferior frontal gyri (IFG; BA 45/47), as shown in Table 2 and Fig. 2. Conjunction of three types of the indirect reply conditions relative to the DR condition identified activation in dmPFC (BA 9/10), bilateral MTG (BA 20/38), bilateral TPJ (BA 39/37), precuneus (BA 17), and left IFG (BA 47), as shown in Table 2 and Fig. 2. A parametric modulation analysis was performed to ensure that these activations were the results of indirectness, which identified activations in dmPFC (BA 9), bilateral MTG (BA 20/21), bilateral TPJ (BA 39), bilateral IFG (BA 38/47) and precuneus (BA 23), also shown in Table 2 and Fig. 2. Clearly, there is a large overlap among the regions in above analyses, suggesting that the brain areas revealed are indeed important for comprehending indirect replies.

The primary goal of this study was to identify the differential neural substrates involved in the comprehension of indirect replies with various contextual relevance. We hence analyzed the differences between the RR, IRC and IRNC conditions. As shown in

Table 2
Activations for contrasts of interest thresholded at 0.001, with AlphaSim correction. All reported coordinates are in the MNI space. BA, Brodmann area.

Region	BA	Coordinates of local maxima			Z(max)	Cluster Size
		x	y	z		
Pooled indirect replies > DR						
L temporo-parietal junction	39	−42	−70	25	5.13	324
Precuneus	17/23	0	−55	25	4.96	321
L middle temporal gyrus	20/21	−63	−34	−8	4.68	267
R temporo-parietal junction	37/39	54	−58	16	4.26	93
R inferior frontal gyrus	38/45	48	29	−14	4.35	58
L inferior frontal gyrus	45/47	−45	32	−14	4.32	53
R middle temporal gyrus	20/21	63	−37	−14	4.02	34
Conjunction of DR vs. indirect reply conditions						
Medial frontal cortex	9/10	−9	50	34	5.15	512
L middle temporal gyrus	20/21	−54	−4	−23	6.11	307
L temporo-parietal junction	39	−39	−67	22	5.22	294
Precuneus	17	−9	−49	37	3.85	215
R middle temporal gyrus	20/38	54	−4	−35	5.53	190
R temporo-parietal junction	37/39	45	−58	19	4.14	122
L inferior frontal gyrus	47	−33	32	−20	4.87	38
Paramatic modulation analysis						
Medial frontal cortex	9	−9	41	49	5.3	554
L middle temporal gyrus	20/21	−60	−16	−17	5.1	261
L temporo-parietal junction	39	−57	−61	28	5.25	237
R middle temporal gyrus	20	66	−31	−14	5.38	112
Precuneus	23	−3	−52	22	4.31	100
R inferior frontal gyrus	38	45	29	−17	4.45	96
L inferior frontal gyrus	38/47	−45	32	−17	4.45	87
L anterior temporal lobe	20	−42	11	−38	4.6	
R anterior temporal lobe	20/21	63	−1	−17	4.37	83
R middle temporal gyrus	21	57	−58	19	4.19	82
R temporo-parietal junction	39	63	−58	31	3.37	
Medial frontal cortex	10	12	65	13	3.98	29
L middle frontal gyrus	9	−36	8	46	3.74	25
RR > IRC						
R middle frontal gyrus	44	51	23	31	3.92	29
R middle temporal gyrus	21	60	−37	−8	4.74	27

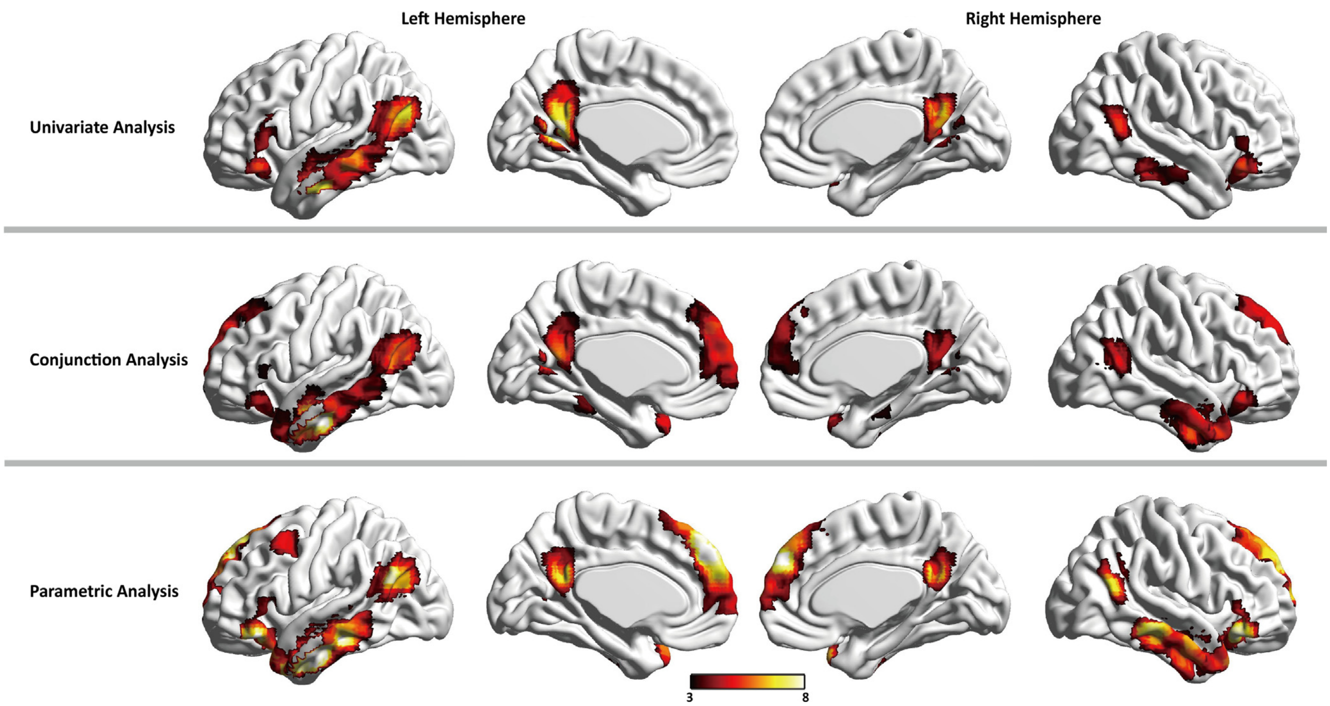

Fig. 2. Activations for univariate analysis (upper panel), conjunction analysis (middle panel) and parametric analysis (lower panel), respectively. All activations are AlphaSim corrected for multiple comparisons. Significant effects are displayed on cortical renderings.

Table 2, the contrast between the RR and IRC conditions showed greater activation in right middle frontal gyrus (MFG; BA 44) and right middle MTG (BA 21). In addition, when we relaxed the voxel-level threshold to $p < 0.005$ uncorrected and a cluster-level threshold of $p < 0.05$ (AlphaSim corrected for multiple comparisons), a main effect of IRC relative to IRNC identified activation in right posterior MTG (BA 21) and left insula (BA 48). Compared with the increased activation in the RR, IRC and IRNC conditions, we found that the activation in right MTG for each indirect replies condition was anterior-posterior gradient. To reveal their differential contributions to the three types of indirect replies, we extracted the parameter estimates from three areas distributed along anterior-to-posterior MTG (MNI coordinates: [54, −4, −35], [60, −37, −8], and [66, −52, 10], see Fig. 3A). We then performed a 3 (position) × 3 (indirect effects, i.e. RR > DR, IRC > DR, and IRNC > DR) repeated-measures ANOVA on the parameter estimates and found a significant position-by-reply type interaction $F(4,88) = 2.82$, $p = 0.030$.

Detailed analyses were conducted to tear apart this interaction (see Fig. 3B). The activation in anterior right MTG slightly increased as contextual relevance decreased (only the indirect effect in RR vs. IRNC was marginally significant, $p = 0.053$). The middle right MTG was sensitive to the literal irrelevance of the dialogue (the indirect effect in both the IRC and IRNC conditions was significantly higher than in the RR condition, $ps < 0.01$, while the indirect effect in the IRC and IRNC condition was not significantly different, $p > 0.9$). The posterior region was sensitive to the absence of contextual hint (indirect effect in the IRNC was significantly higher than that in both the RR and IRC conditions, $ps < 0.01$, while the indirect effect in the RR and IRC conditions did not differ significantly, $p > 0.9$). All p-values in the above analyses were adjusted with Bonferroni correction for multiple comparisons.

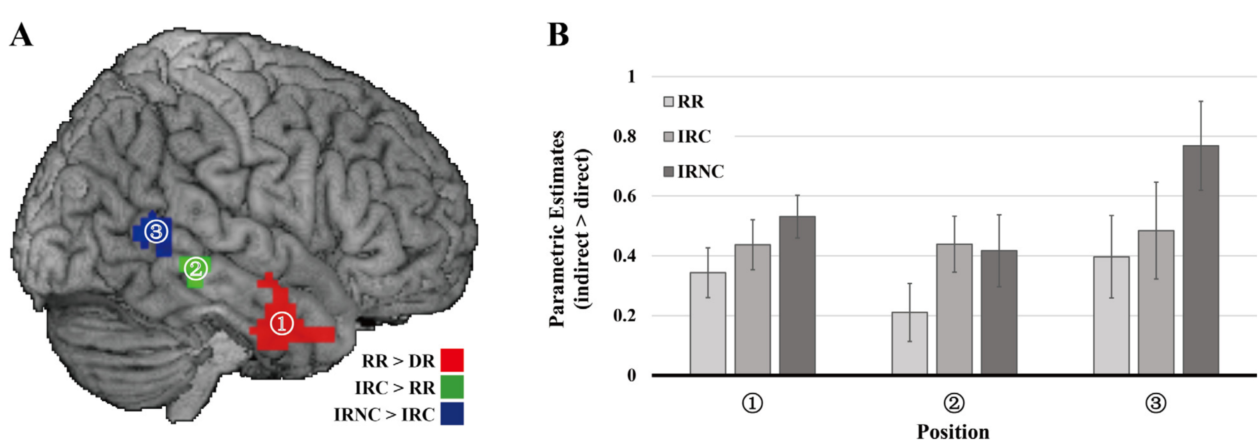

Fig. 3. Anterior-posterior gradient in right MTG. (A) Activations in the right MTG for the RR > DR contrast (red), the IRC > RR contrast (green), the IRNC > IRC contrast (blue), as described in the *Results* Section 3.2. (B) Plot of the indirect effects in the right MTG. The error bars depict the standard errors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

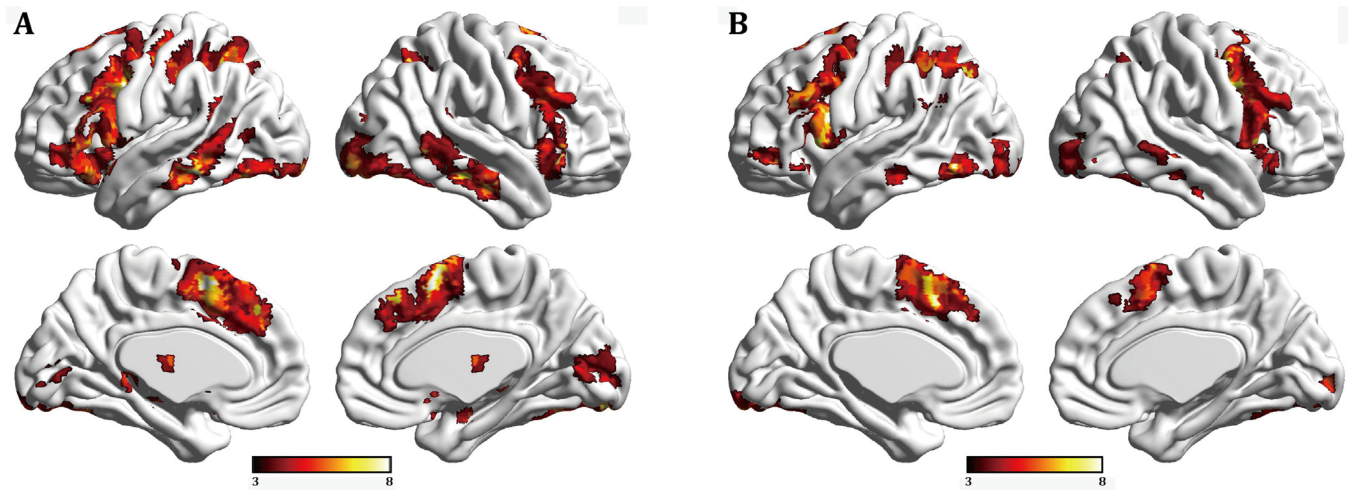

Fig. 4. Results of psychophysiological interaction (PPI) analysis using right TPJ as the source region (A) and dmPFC as the source region (B). During comprehending the (pooled) indirect replies (vs. direct replies), increased connectivity with right TPJ seed was observed in Panel A, and that with dmPFC seed was in Panel B. All PPI activations are AlphaSim corrected for multiple comparisons. Significant effects are displayed on cortical renderings.

3.3. Psychophysiological interaction (PPI)

We conducted PPI analysis to find brain regions in which functional connectivity with right TPJ and dmPFC (from the pooled indirect replies vs. DR contrast) was modulated by the generation of CI. As shown in Fig. 4A, right TPJ showed increased functional connectivity with bilateral middle/superior temporal gyri (BA 21/22), bilateral IFG (BA 44/45/47), precentral/postcentral gyrus, supplementary motor area (SMA; BA 6), inferior parietal lobe (IPL), dorsal striatum (caudate nucleus and putamen), bilateral thalamus, and bilateral middle/inferior occipital gyri for the indirect replies (relative to the direct replies). As shown in Fig. 4B, dmPFC showed increased functional connectivity with bilateral middle temporal gyri (BA 21/22), bilateral IFG (BA 44/45/47), SMA (BA 6), bilateral TPJ (BA 40) extending to IPL, dorsal striatum, and bilateral middle/inferior occipital gyri for the indirect replies (relative to the direct replies).

4. Discussion

By comparing direct and indirect replies, this study aimed to identify brain regions involved in comprehending CI during conversation. The difference in brain activation between direct and indirect replies reflects the generation of implicated meanings since the replies we used were literally the same across all four conditions (the direct and the three indirect conditions). Results of univariate analysis, conjunction analysis and parametric analysis consistently showed that understanding indirect replies activated both the frontal-temporal language network, including bilateral IFG and MTG, and ToM-related brain areas, such as rTPJ, dmPFC and precuneus. Our results replicated in Chinese the essential findings of previous studies on indirect speech (Bašnáková et al., 2014, 2015; Shibata et al., 2011; van Ackeren et al., 2012, 2016).

As hypothesized, bilateral IFG and MTG of the language network were recruited by CI processing. Although IFG was found to be related to grammatical categorization (Ni et al., 2000), syntax parsing (Friederici & Kotz, 2003), and pronoun resolution (Nieuwland, Petersson, & Van Berkum, 2007), the enhanced activation in bilateral IFG here was recruited to process semantic or high-level pragmatic information during understanding indirect replies, since confounding variables were well controlled here. On the semantic aspect, IFG was found to be associated with semantic information retrieval (Wagner, Paré-Blagoev, Clark, & Poldrack, 2001), semantic unification (Hagoort, Hald, Bastiaansen, & Petersson, 2004; Zhu et al., 2012), and selection (Kan & Thompson-Schill, 2004; Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997). More specifically, the activation in bilateral IFG might reflect higher cognitive demands of unifying semantic information into context during non-literal language comprehension (Menenti et al., 2009; Rapp et al., 2004, 2007, 2012). Moreover, MTG is considered as a core region of the semantic system (Binder, Desai, Graves, & Conant, 2009; Visser & Lambon Ralph, 2011; Xu, Kemeny, Park, Frattali, & Braun, 2005). During pragmatic processing, these two areas are responsible for interpreting text or narrative information (Ferstl & von Cramon, 2001; Fletcher et al., 1995; Xu et al., 2005), and understanding social interaction information (Ross & Olson, 2010; Zahn et al., 2007). According to pragmatics accounts, addressees would treat indirect speech as a linguistic expression that is semantically related to the context. In fact, addressees assume that all utterances should yield to relevance principles; they try to find the optimal relevance between the utterance and its context and to complete propositional representation (Bach, 1994; Carston, 2004). Hence, the drive to recover the optimal interpretation of indirect speech requires greater participation of semantic activation, selection and integration.

More interestingly, our findings provided a specific profile of the core language system in processing indirect reply. By contrasting the three types of indirect replies with the region of interest analysis, we showed that right MTG responded differentially to the changes in contextual relevance of the utterance. As literal relevance between the reply and its context weakens, anterior-posterior gradient in right MTG is activated: the anterior region of right MTG is increasingly activated when the reply is indirect; the middle region is increasingly activated when literal relevance between the utterance and its immediate context (question) is absent; and right posterior MTG is increasingly activated only when out-context knowledge should be involved in the processing of pragmatic inference. These results are in line with several studies showing a strong involvement of right hemisphere in comprehending contextual and figurative meaning and in multi-sentence and discourse level processing (for a review, see Bookheimer, 2002). As opposed to left hemisphere, right hemisphere is more sensitive to discourse modulation, such as textual coherence and discourse anomalies (Kuperberg et al., 2006; Menenti et al., 2009; Nieuwland, 2012). In particular, studies on irony, metaphor and idiom also observed the engagement of right MTG (for meta-analysis see Rapp et al., 2012), which indicates that right MTG plays a role in processing ambiguous or incoherent discourses. According to the fine-coarse coding hypothesis (Beeman & Chiarello, 1998), the right hemisphere performs relatively coarser semantic coding, which allows efficient semantic integration of all accessible meanings within a broad semantic field. On the contrary, the left hemisphere performs fine semantic coding and the processing in left MTG is supported by the coarse-coding in right MTG. In particular, right anterior MTG supports the processing of semantic integration while right posterior MTG supports weak and diffuse semantic activation (for a review see Jung-Beeman, 2005). In the current study, the increased engagement of the middle and posterior regions in right MTG suggests that relatively distant semantic information is activated and unified in order to construct the optimal contextual relevance. Just as theoretical accounts suggested (e.g. Bach, 1994; Borg, 2009; Sperber & Wilson, 1986), the generation of CI is supported by semantic activation, selection and integration, and thus relevant contextual features and additional conceptual representations can be brought into the semantic field.

It has been shown that extra-language areas are involved in understanding nonliteral meaning (Bašnáková et al., 2014; Jang et al., 2013; Shibata et al., 2011; van Ackeren et al., 2012). In this study, indirect reply also elicited enhanced activations in TPJ, dmPFC, and precuneus, relative to direct reply. TPJ, which refers to the junction of temporal lobe and parietal lobe (Carter & Huettel, 2013), is considered as a type of 'nexus' that supports the extraction and integration of social contexts for behavior (Carter, Bowling, Reeck, & Huettel, 2012; Carter & Huettel, 2013; Schaafsma, Pfaff, Spunt, & Adolphs, 2015). DmPFC is also found to be associated with inductive reasoning (Siebörger et al., 2007) and theory-of-mind processes (Fletcher et al., 1995). As our activation pattern in bilateral TPJ, dmPFC and precuneus is typical for ToM processing (Koster-Hale & Saxe, 2013; Mar, 2011), we suggest that this set of activations is recruited by a ToM-like inferential processing. In other words, to understand indirect speech, listeners need to infer the speaker's aim and intention, which requires higher-order ToM-like mentalizing.

PPI analyses revealed that both rTPJ and dmPFC showed significantly higher functional coupling with bilateral IFG and MTG during the processing of indirect replies, relatively to the processing of direct replies. rTPJ and dmPFC are the most representative regions of ToM network; the neural activity in which reflects the processing of mentalizing other people's beliefs and intentions (Koster-Hale & Saxe, 2013; Saxe, 2003). According to previous studies, bilateral IFG and MTG, as described above, are critical parts of the fronto-temporal semantic network. Thus, our findings not only agree with recent studies in that the increased neural interaction of left (and right) IFG and mPFC supports the processing of pragmatic inference (Spotorno et al., 2012; van Ackeren et al., 2016), but also reveal the global communication between the core semantic network (including IFG and MTG) and the ToM network (including TPJ and mPFC) in understanding indirect replies. In addition, we observed significantly increased functional connectivity between rTPJ/dmPFC and other brain regions, such as SMA, IPL, dorsal striatum, and thalamus. Task difficulty of indirect reply conditions is generally higher than that of direct reply condition as it requires additional inferential processing. To meet the task requirement for the indirect reply conditions, participants had to spend greater cognitive resources. Indeed, several neuroimaging studies have shown that the brain regions listed above are recruited in the processing of cognitive control (Bush & Shin, 2006; Corbetta & Shulman, 2002; Nagano-Saito, Martinu, & Monchi, 2014). It is possible that this increased connectivity reflects the domain-general cognitive processing of completing a difficult task.

5. Conclusion

To conclude, we presented evidence suggesting that comprehending conversational implicatures requires complex semantic and pragmatic processing. In particular, when an utterance is literally less relevant to its context, relatively coarser and broader semantic information needs to be activated, subserved in part by right MTG in an anterior-posterior gradient manner, and integrated to construct optimal contextual relevance. During pragmatic inference, ToM-like inferential processes used for recovering the speaker's meaning relies on both the literal meaning of the utterance and the richer semantic/pragmatic information derived from specific context.

Acknowledgements

The authors thank Dr. Weidong Zhan, Dr. Sujian Li, and Mr. Liang Wang for help with Latent Semantic Analysis, and Mr. Wenshuo Chang for help with the collection of scenario materials. The authors also thank the two anonymous reviewers for their constructive comments on an earlier version of the manuscript. This study was supported by grants from the Natural Science Foundation of China [31470976] and the Social Science Foundation of China [12&ZD119] awarded to Prof. Xiaolin Zhou.

References

Ashbumer, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage, 26*(3), 839–851. http://dx.doi.org/10.1016/j.neuroimage.2005.02.018.

Bach, K. (1994). Conversational impliciture. *Mind & Language, 9*(2), 124–162. http://dx.doi.org/10.1111/j.1468-0017.1994.tb00220.x.

Bach, K. (2006). The Top 10 Misconceptions about Implicature. In B. J. Birner & G. L. Ward (Eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn* (pp. 21–30). Amsterdam: John Benjamins Publishing. http://dx.doi.org/10.1075/slcs.80.03bac.

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. http://dx.doi.org/10.1016/j.neuropsychologia.2015.03.030.

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the Language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex, 24*(10), 2572–2578. http://dx.doi.org/10.1093/cercor/bht112.

Beeman, J. M., & Chiarello, C. (1998). Complementary right and left hemisphere language comprehension. *Current Directions in Psychological Science, 7*(1), 2–8. http://dx.doi.org/10.1111/1467-8721.ep11521805.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*(12), 2767–2796. http://dx.doi.org/10.1093/cercor/bhp055.

Bohrn, I. C., Altmann, U., & Jacobs, A. M. (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. http://dx.doi.org/10.1016/j.neuropsychologia.2012.07.021.

Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience, 25*(1), 151–188. http://dx.doi.org/10.1146/annurev.neuro.25.112701.142946.

Borg, E. (2009). On Three Theories of Implicature: Default theory, relevance theory and minimalism. *International Review of Pragmatics, 1*(1), 63–83. http://dx.doi.org/10.1163/187731009X455848.

Bush, G., & Shin, L. M. (2006). The multi-source interference task: An fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nature Protocols, 1*(1), 308–313. http://dx.doi.org/10.1038/nprot.2006.48.

Carston, R. (2004). Relevance theory and the saying/implicating distinction. In L. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 633–656). Oxford: Blackwell. http://dx.doi.org/10.1002/9780470756959.ch28.

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science, 337*(6090), 109–111. http://dx.doi.org/10.1126/science.1219681.

Carter, R. M., & Huettel, S. A. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences, 17*(7), 328–336. http://dx.doi.org/10.1016/j.tics.2013.05.007.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215. http://dx.doi.org/10.1038/nrn755.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for*

Information Science, 41(6), 391–2237. http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9.

Egorova, N., Pulvermüller, F., & Shtyrov, Y. (2014). Neural dynamics of speech act comprehension: An MEG study of naming and requesting. *Brain Topography, 27*(3), 375–392. http://dx.doi.org/10.1007/s10548-013-0329-3.

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping, 29*(5), 581–593. http://dx.doi.org/10.1002/hbm.20422.

Ferstl, E. C., & von Cramon, D. Y. (2001). The role of coherence and cohesion in text comprehension: An event-related fMRI study. *Cognitive Brain Research, 11*(3), 325–340. http://dx.doi.org/10.1016/S0926-6410(01)00007-6.

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

Friederici, A. D., & Kotz, S. A. (2003). The brain basis of syntactic processes: Functional imaging and lesion studies. *Neuroimage, 20*(Suppl. 1), S8–S17. http://dx.doi.org/10.1016/j.neuroimage.2003.09.003.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage, 6*(3), 218–229. http://dx.doi.org/10.1006/nimg.1997.0291.

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., & Worsley, K. J. (1999). Multisubject fMRI studies and conjunction analyses. *Neuroimage, 10*(4), 385–396. http://dx.doi.org/10.1006/nimg.1999.0484.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics III: Speech acts* (pp. 41–58). New York: Academic Press.

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology, 4*, 416. http://dx.doi.org/10.3389/fpsyg.2013.00416.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438–441. http://dx.doi.org/10.1126/science.1095455.

Hagoort, P., & Levinson, S. C. (2014). Neuropragmatics. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 667–674). Cambridge, MA: MIT Press.

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

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences, 9*(11), 512–518. http://dx.doi.org/10.1016/j.tics.2005.09.009.

Kan, I. P., & Thompson-Schill, S. L. (2004). Selection from perceptual and conceptual representations. *Cognitive, Affective, & Behavioral Neuroscience, 4*(4), 466–482. http://dx.doi.org/10.3758/CABN.4.4.466.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848. http://dx.doi.org/10.1016/j.neuron.2013.08.020.

Kuperberg, G. R., Lakshmanan, B. M., Caplan, D. N., & Holcomb, P. J. (2006). Making sense of discourse: An fMRI study of causal inferencing across sentences. *Neuroimage, 33*(1), 343–361. http://dx.doi.org/10.1016/j.neuroimage.2006.06.001.

Ledberg, A., Åkerman, S., & Roland, P. E. (1998). Estimation of the probabilities of 3D clusters in functional brain images. *Neuroimage, 8*(2), 113–128. http://dx.doi.org/10.1006/nimg.1998.0336.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual review of psychology, 62*, 103–134. http://dx.doi.org/10.1146/annurev-psych-120709-145406.

Menenti, L., Petersson, K. M., Scheeringa, R., & Hagoort, P. (2009). When elephants fly: Differential sensitivity of right and left inferior frontal gyri to discourse and world knowledge. *Journal of Cognitive Neuroscience, 21*(12), 2358–2368. http://dx.doi.org/10.1162/jocn.2008.21163.

Nagano-Saito, A., Martinu, K., & Monchi, O. (2014). Function of basal ganglia in bridging cognitive and motor modules to perform an action. *Frontiers in Neuroscience, 8*, 1–12. http://dx.doi.org/10.3389/fnins.2014.00187.

Ni, W., Constable, R. T., Mencl, W. E., Pugh, K. R., Fulbright, R. K., Shaywitz, S. E., ... Shankweiler, D. (2000). An event-related neuroimaging study distinguishing form and content in sentence processing. *Journal of Cognitive Neuroscience, 12*(1), 120–133. http://dx.doi.org/10.1162/08989290051137648.

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage, 25*(3), 653–660. http://dx.doi.org/10.1016/j.neuroimage.2004.12.005.

Nieuwland, M. S. (2012). Establishing propositional truth-value in counterfactual and real-world contexts during sentence comprehension: Differential sensitivity of the left and right inferior frontal gyri. *Neuroimage, 59*(4), 3433–3440. http://dx.doi.org/10.1016/j.neuroimage.2011.11.018.

Nieuwland, M. S., Petersson, K. M., & Van Berkum, J. J. (2007). On sense and reference: Examining the functional neuroanatomy of referential processing. *Neuroimage, 37*(3), 993–1004. http://dx.doi.org/10.1016/j.neuroimage.2007.05.048.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526. http://dx.doi.org/10.1017/S0140525X00076512.

Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., & Kircher, T. T. (2004). Neural correlates of metaphor processing. *Cognitive Brain Research, 20*(3), 395–402. http://dx.doi.org/10.1016/j.cogbrainres.2004.03.017.

Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., & Kircher, T. T. (2007). Laterality in metaphor processing: Lack of evidence from functional magnetic resonance imaging for the right hemisphere theory. *Brain and Language, 100*(2), 142–149. http://dx.doi.org/10.1016/j.bandl.2006.04.004.

Rapp, A. M., Mutschler, D. E., & 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. http://dx.doi.org/10.1016/j.neuroimage.2012.06.022.

Ross, L. A., & Olson, I. R. (2010). Social cognition and the anterior temporal lobes. *Neuroimage, 49*(4), 3452–3462. http://dx.doi.org/10.1016/j.neuroimage.2009.11.012.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72. http://dx.doi.org/10.1016/j.tics.2014.11.007.

Shibata, M., Abe, J.-I., 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. http://dx.doi.org/10.1016/j.neuropsychologia.2011.09.006.

Siebörger, F. T., Ferstl, E. C., & von Cramon, D. Y. (2007). Making sense of nonsense: An fMRI study of task induced inference processes during discourse comprehension. *Brain Research, 1166*, 77–91. http://dx.doi.org/10.1016/j.brainres.2007.05.079.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition* (Vol. 1) Cambridge, MA: Harvard University Press.

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

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

Uchiyama, H. T., Saito, D. N., Tanabe, H. C., Harada, T., Seki, A., Ohno, K., ... Sadato, N. (2012). Distinction between the literal and intended meanings of sentences: A functional magnetic resonance imaging study of metaphor and sarcasm. *Cortex, 48*(5), 563–583. http://dx.doi.org/10.1016/j.cortex.2011.01.004.

van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Ruesschemeyer, 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. http://dx.doi.org/10.1162/jocn_a_00274.

van Ackeren, M. J., Smaragdi, A., & Rueschemeyer, S.-A. (2016). Neuronal interactions between mentalizing and action systems during indirect request processing. *Social Cognitive and Affective Neuroscience, 11*(9), 1402–1410. http://dx.doi.org/10.1093/scan/nsw062.

Visser, M., & Lambon Ralph, M. A. (2011). Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of Cognitive Neuroscience, 23*(10), 3121–3131. http://dx.doi.org/10.1162/jocn_a_00007.

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron, 31*(2), 329–338. http://dx.doi.org/10.1016/S0896-6273(01)00359-2.

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *Neuroimage, 25*(3), 1002–1015. http://dx.doi.org/10.1016/j.neuroimage.2004.12.013.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences, 104*(15), 6430–6435. http://dx.doi.org/10.1073/pnas.0607061104.

Zhu, Z., Hagoort, P., Zhang, J. X., Feng, G., Chen, H. C., Bastiaansen, M., & Wang, S. (2012). The anterior left inferior frontal gyrus contributes to semantic unification. *Neuroimage, 60*(4), 2230–2237. http://dx.doi.org/10.1016/j.neuroimage.2012.02.036.