Involvement of left inferior frontal gyrus in sentence-level semantic integration

Zude Zhu a, John X. Zhang b, Suiping Wang a, Zhuangwei Xiao c, Jian Huang a, Hsuan-Chih Chen b,⁎

a Department of Psychology, South China Normal University, Guangzhou, China
b Department of Psychology, Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China
c Guangdong Province Key Laboratory of Medical Molecular Imaging, Medical School of Shantou University, Shantou, China

A R T I C L E  I N F O

Article history:
Received 14 October 2008
Revised 10 April 2009
Accepted 28 April 2009
Available online 6 May 2009

Keywords:
Sentence comprehension
Semantic integration
Left inferior frontal gyrus (LIFG)
Violation paradigm
fMRI

A B S T R A C T

Using event-related functional MRI, we examined the involvement of the left inferior frontal gyrus (LIFG) in semantic integration in reading Chinese sentences. During scanning, Chinese readers read individually presented sentences and judged whether or not a sentence was semantically acceptable. Behaviorally, those sentences with a small degree of semantic violation were found to be more difficult to reject relative to sentences with a large degree of semantic violation, indicating that more semantic integration occurred in the former than in the latter condition. Direct contrast revealed significantly greater brain activity in the LIFG for sentences with a small violation, relative to those with a large violation, but no differences in any anterior temporal cortical areas between the two types of anomalous sentences. The results are in line with the idea that the LIFG plays a critical role in integrating individual word meanings to coherent sentence-level messages, but not with the idea that semantic integration depends on anterior temporal cortex in language comprehension.

© 2009 Elsevier Inc. All rights reserved.

Introduction

In recent years, there has been an increasing interest in language processing and its neural correlates (e.g., Constable et al., 2004; Hagoort, 2003, 2005; Hickok and Poeppel, 2004; Jung-Beeman, 2005; Spitsyna et al., 2006; Wang et al., 2008; Willems et al., 2008). As a result, different models have been postulated on the basis of the findings in cognitive neuroscience research.

A major model on the topic is the Bilateral Activation, Integration, and Selection (BAIS) model, which proposes that semantic processing in language comprehension consists of three distinct but highly interactive components (i.e., activation, selection, and integration), each of which is supported by different brain areas (Jung-Beeman, 2005). Specifically, bilateral Wernicke’s areas are considered to be responsible for activating or retrieving lexical and semantic information, the inferior frontal gyrus (IFG) for selection among competing semantic activations, and anterior temporal areas for semantic integration to compute semantic overlap among multiple semantic activations and to construct higher-order semantic relations.

Another major model on the same topic is the one proposed by Hagoort and colleagues (Hagoort, 2003, 2005; Hagoort and van Berkum, 2007), in which they argue that a different set of three major components (i.e., memory, unification, and control) is crucial for language comprehension and production. According to this Memory, Unification, and Control (MUC) model, the left temporal cortex underlies the memory component in retrieving word information from long-term memory, the LIFG underlies the unification component in combining retrieved information into larger units, and the anterior cingulate cortex and dorsolateral prefrontal cortex (DLPFC) underlie the control component in planning verbal actions and allocating attention.

Semantic integration is a key component in the construction of sentence-level meaning representations, but different researchers attribute this process to different brain areas (i.e., anterior temporal cortex vs. LIFG). To examine which one of the two mentioned cortical areas is connected with semantic integration in language processing, the present study adopted a violation paradigm commonly used for studying semantic processing (e.g., Dapretto and Bookheimer, 1999; Hagoort et al., 2004; Kutas and Hillyard, 1980; see review in Kaan and Swaab (2002) and Kutas et al. (2006)). The violation paradigm typically involves comparison of normal sentences with anomalous sentences that have been violated in a certain manner (e.g., a semantic violation in the present study). The basic assumption is that participants are subject to increasing demand on and recruit more cognitive resources for integration in reading anomalous sentences. Following the same logic, brain regions more strongly activated by the anomalous sentences relative to the normal sentences are generally interpreted as possible neural correlates of a particular type of reading related operations (e.g., Baumgaertner et al., 2002; Hagoort et al., 2004; Homae et al., 2002).

It is important to note, however, that there are two potential problems in using the above mentioned violation paradigm to investigate semantic integration in language processing. First, normal sentences and semantically violated sentences differ not only in...
semantic integration, but also in other aspects. For example, anomalous sentences may recruit processes not engaged by normal sentences at all, such as violation detection and violation repairing (Indefrey et al., 2001; Kaan and Swaab, 2003). These processes, though unrelated to semantic integration, may also show up in the anomalous vs. normal comparison. Although similar problems have been recognized in previous syntactic violation studies (e.g., Indefrey et al., 2001), no attempt has been made so far to deal with it in semantic studies.

Another potential problem is that, when comparing the anomalous and normal sentences, the two types of sentences are actually associated with different responses. In fact, in making semantically acceptable or unacceptable judgments, participants are essentially matching the constructed sentence meaning with some previously established, semantic representations so that semantically acceptable sentences require a match or positive response and semantically unacceptable sentences require a mismatch or negative response. It has been well established in cognitive psychology that positive and negative decisions may involve different cognitive processes (see, e.g., Farell (1985), Sternberg (1966) and Treisman and Gormican (1988) for evidence in behavioral studies; e.g., Zhang et al. (2003) for evidence in brain imagining studies), and that making positive and negative responses may be influenced by factors such as task demand, response bias, and response strategy (see review in Farell (1985)). Hence, comparison of anomalous and normal sentences may reflect general decision-making processes not specific to semantic integration. In other words, semantic integration is only one of several factors that may differentiate anomalous and normal conditions, if the two conditions require different responses. Under this situation, one should be cautious in interpreting brain activation differences across the two conditions.

In the present study, we attempted to make use of the principle of parametric design to compare two sentence types with different degrees of semantic violation. Table 1 illustrates the experimental conditions, each with a sample sentence. The anomalous sentences were constructed by changing a target noun in a normal sentence. For example, when the normal sentence is as follows: “The construction workers use pump to draw underground water” (English translation), the target word “pump” is replaced with “iron” to produce a relatively small semantic violation, and with “salt” to produce a relatively large semantic violation.

Although neither “iron” nor “salt” can be integrated with the rest of the sentence to produce a coherent meaning representation, the former word is semantically related to the original word “pump”, and is more likely to fit the sentence context, relative to the latter, which is semantically unrelated to the original word (see similar manipulations in Federmeier and Kutas (1999)). Consequently, we expected that it would be more difficult for participants to detect the violation in a small semantically violated sentence and reject the sentence as semantically acceptable, relative to a large semantically violated sentence. This is because participants would need to engage more semantic integration in order to make sense of the sentence in the small violation condition than in the large violation condition.

Relative to the anomalous vs. normal contrast, the small vs. large violation contrast would be much more comparable, because they involve the same processes of violation detection, semantic repair, and require the same negative response. Therefore, the contrast between the small vs. large violation should be more appropriate to reveal the neural correlates of semantic integration and to test the specific predictions from models such as the MUC and the BAIS.

Method

Participants

Sixteen right-handed native Chinese speakers (8 males and 8 females, aged between 20 and 27 years, with a mean age of 23.7 years) participated in this study. All had normal or corrected-to-normal vision and none had any history of psychiatric or neurological disorders. Written informed consent was obtained from each participant following a protocol approved by the IRB of the Medical College of Shantou University, Shantou, China.

Stimuli

The stimuli consisted of three types of sentences, including normal sentences, sentences with a small semantic violation, and those with a large semantic violation. The normal sentences were 66 grammatically correct and meaningful Chinese sentences. Each normal sentence was modified by replacing a target word (e.g., 'pump') with a ‘related’ word (e.g., ‘iron’) to construct a small violation sentence and an ‘unrelated’ word (e.g., ‘salt’) to construct a large violation sentence, as shown in Table 1.

The ‘related’ word was chosen from a semantic category related to the original target word. The ‘unrelated’ word, in contrast, was chosen from a category unrelated to the original target word and was therefore unlikely to be related in an explicit manner to the target word.

The effectiveness of such stimulus manipulation was confirmed in a pilot rating study using a 5-point scale with a separate group of 10 participants (from the same subject pool). The scale was so labeled that 1 stands for “minimal relationship”, 2 for “weak relationship”, 3 for “medium relationship”, 4 for “strong relationship”, and 5 for “very strong relationship”. The average rating for semantic relatedness between the original target words and the ‘related’ words was 2.81 (SD = 0.46), and that between the target words and the ‘unrelated’ words was 1.36 (SD = 0.15). The two values were significantly different from each other (paired t-test, t (9) = 12.59, p < 0.001).

Table 1

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sample sentence</th>
<th>RT (ms)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>施工人员用 水泵 抽取 地下水。&lt;br/&gt;/shigongrenyuan yong shuibeng chouqu dixiashui/</td>
<td>2820 (135)</td>
<td>90.9 (2.3)</td>
</tr>
<tr>
<td>Small violation</td>
<td>施工人员用 钢铁 抽取 地下水。&lt;br/&gt;/_gangtie/ _&lt;br/&gt;/iron/ _&lt;br/&gt;2575 (139)</td>
<td>92.1 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Large violation</td>
<td>施工人员用 食盐 抽取 地下水。&lt;br/&gt;/_shiyan/ _&lt;br/&gt;/salt/ _&lt;br/&gt;2346 (137)</td>
<td>97.7 (10)</td>
<td></td>
</tr>
</tbody>
</table>

Target words are underlined. SEs are shown in brackets.
All the target words used were concrete two-character noun words and therefore matched in concreteness and word length. Besides, they were high-frequency words (Liu et al., 1990). The average occurrences per million for the normal, small, and large violation sentences were 42.4 (SD = 37.9), 41.1 (SD = 36.5), and 42.9 (SD = 39.2), respectively, with no significant differences across sentence types ($F (2, 130) = 1.5, p = 0.22$). The visual complexity of the two characters in the target word was also matched in terms of the average stoke numbers, which were 8.27 ± 0.29, 8.57 ± 0.29, and 8.41 ± 0.30 (mean ± SE) for the congruous sentence, the small violation sentences, the large violation sentences, respectively, with no significant differences across conditions ($p = 0.5$). In a 5-point imaginability scale (1 for high and 5 for low), a group of 16 college students rated the target words and gave a mean value of 1.28 ± 0.06, 1.33 ± 0.06, 1.34 ± 0.08, for the congruous sentences, the small violation sentences, and the large violation sentences respectively, with no significant differences across conditions ($p = 0.2$).

Given the flexibility of word order in Chinese grammar (Chen, 1999; 2006; Chen et al., 2000), early local anomalies in a sentence that prompt a clear violation in similar English sentences may end up being perfectly interpretable and meaningful in Chinese. For example, even though “eat table” does not seem to be acceptable, sentences such as 小花 [Xiao Ming] 吃了 (eat) 菜 (table) 上的 (on) 苹果 (apple)” (translated as, Xiao Ming eats the apples on the table) are common and fully acceptable. Recognizing this fact, we did not put all target words at the end of a sentence so that we could include a variety of sentences for greater generality.

In addition to the sentence structure shown in Table 1 involving an agent performing an action for a specific goal, there were two other types of structures, one being descriptions of real world situations or scenarios (e.g., 刘洁的服装店里挂满了裤子/眼镜/坟墓 translated as, In the fashion store of Liu jie there hang many trousers/glasses/tombs), and one being descriptions of common facts (e.g., 果汁/荔枝/舞蹈是一种经常见到的饮料, translated as, Juice/lychee/dancing is a commonly seen drink). This variation in sentence structure was to increase ecological validity and to prevent participants from adopting specific response strategies if sentence structure was fixed.

To enhance ecological validity, the cloze probability of the sentence context was varied. This would also help to reduce anticipation and adoption of specific response strategies. Thirty additional participants from the same subject pool were given all the sentences with the target word replaced by a blank and asked to fill in the blank with the first one that came to mind and made the sentence meaningful. For each sentence, the frequency of the word given by the most number of participants, scaled by the total participant number (i.e., 30), was taken as the cloze probability for that sentence. The mean probability averaged over all sentences was 41.9% with a standard deviation of 19.0. This value was lower than the 58.8% defined in Federman and Kutas (1999) as “low constraint”, suggesting that for our materials, it would be difficult for participants to anticipate the target words based on the sentence context.

A separate group of 48 participants from the same subject pool rated the semantic plausibility of all sentences on a 7-point scale, with 1 for “extremely unacceptable” and 7 for “totally acceptable”. The average ratings for the normal, small, and large violation sentences were 5.6 (SD = 0.6), 2.3 (SD = 0.7), and 1.7 (SD = 0.6), respectively. There were significant differences in rating scores across sentence type ($F (2, 94) = 679.2, p < 0.001$), validating our manipulation of semantic acceptability in the three types of sentences. The rating for the normal sentences was significantly higher than that for either type of the violation sentences, and the rating for the small violation condition was significantly higher than that for the large violation sentences (normal vs. small violation, $t (47) = 24.7$; normal vs. large violation, $t (47) = 29.0$; small violation vs. large violation, $t (47) = 9.2$; all corrected $p < 0.001$).

**Procedure**

The E-Prime software package (Psychology Software Tools, Pittsburgh, PA) was used for stimulus presentation and response collection. Before getting into the scanner, participants completed a practice session of 12 trials. The practice was repeated until they were accurate in more than 10 trials. The experimenter discussed their errors with them to make sure they understood the task. The first session inside the scanner was also for practice (i.e., for participants to adjust to the general setup inside the scanner). Thus, the participants were familiarized with the task and the scanning environment. They then completed four test runs inside the scanner, each lasting 5 min and 42 s. Each run included a section of 22 trials, preceded and followed by a 30-s passive viewing of a fixation cross to enhance estimation of baseline signals. In each trial, all words of a sentence were centrally displayed at the same time and stayed on screen until a response was made. The participants were asked to judge whether each presented sentence was semantically acceptable or not, and to indicate their response with a left or right key press. Both response time (RT) and accuracy were recorded.

As in our previous studies (Wang et al., 2008; Xiao et al., 2005), the present design was a rapid event-related design where all types of sentences were randomly intermixed in each run. The intervals between sentence onsets in two consecutive trials were pseudo-randomized to vary from 10 to 16 s with a mean value of 13 s. Each participant read all of the 66 sentences, 22 in the normal version, 22 in the small violation version, and 22 in the large violation version. With proper counterbalancing, each sentence appeared in its three versions equally likely across all participants. There were 22 additional normal sentences, which were included as fillers to balance the positive and negative responses. Each sentence type was randomly distributed in the four test runs with approximately equal probability.

**Image acquisition**

The scanning was conducted on a 1.5 T Philips MRI scanner at the Medical College of Shantou University, using a standard headcoil. Twenty axial slices covering the whole brain were acquired with a T2*-weighted gradient-echo planar imaging (EPI) pulse sequence (TR = 2 s, TE = 45 ms, flip angle = 90°, FOV = 23 cm × 23 cm, matrix = 64 × 64, slice thickness = 6 mm, no gap). Coplanar anatomical images were acquired with a T1-weighted spin echo pulse sequence (matrix = 256 × 256, TR = 204 ms, TE = 14 ms) for co-registration. A high-resolution 3D volume was also acquired with spoiled gradient recalled echo (SPGR) sequence for spatial normalization. Participants lay supine inside the scanner, wearing MRI-compatible earphones and goggles (Resonance Technology Company, Los Angeles, CA) and holding a button box. They were told not to move their head (restrained in padding) inside the scanner and that they could close their eyes for a short rest between two successive runs.

**Data analysis**

The behavioral data were analyzed with a one-way repeated-measure analysis of variance (ANOVA) separately for response time and accuracy, with sentence type as the factor (three levels: normal, small violation, large violation sentences), followed by post-hoc analysis performed on each of the three pair-wise comparisons with corrections for multiple comparisons.

The imaging data were analyzed with the AFNI software package (Cox, 1996), following a procedure similar to our previous work (Wang et al., 2008). Preprocessing of the functional datasets included linear drifts, outliers, and motion artifacts removal, smoothing (FWHM = 6, Gaussian kernel), spatial normalization, and transformation to Talairach space (Talairach and Tournoux, 1988) with 3 mm resolution. Different sentence types were modeled with different trial
events. Hemodynamic responses for each event were estimated for each individual participant with the 3Ddeconvolve program, using the six motion correction parameters from preprocessing (x, y, z translations, roll, pitch, and yaw) as covariates to remove residual motion artifacts. The onset time of each sentence was used to mark the start of this sentence presentation event in the data analysis for the even-related design (Wang et al., 2008).

The estimated percent signal changes, one for each of the three sentence types in each voxel for each participant, were then submitted to the 3dANOVA2 program for a voxel-wise random effect group analysis. This analysis produced four activation maps, three based on t-statistic showing the voxels whose BOLD activity was significantly greater than the resting baseline, one for each of the three sentence type conditions, and the forth based on F statistic showing the voxels whose BOLD activity was significantly modulated by sentence type.

This produced activation maps for each voxel showing activation for each sentence type relative to rest, as well as an activation map with one F statistic indicating the extent to which the brain activity in that voxel was modulated by sentence type.

Significant activations were all with a minimal cluster size of 386 mm$^3$ for a corrected $p<0.05$ threshold, determined with Monte Carlo simulation. Finally, three mean percent signal changes, one for each condition, were computed for every participant in each of the identified activation region, and used for post-hoc pair-wise comparisons across the three sentence type conditions.

Results

Behavioral results

There was a significant effect of sentence type for both response time ($F(2, 30)=32.46$, $p<0.001$) and accuracy ($F(2, 30)=6.20$, $p<0.01$). The average reaction times and accuracy are shown in Table 1. The normal sentences ($M=2820$ ms) produced longer reaction times than did the two types of anomalous sentences ($p<0.05$). In addition, those sentences with a small violation produced longer reaction times than did the sentences with a large violation (2575 ms vs. 2346 ms, $p<0.05$). The mean accuracies were comparable in the normal and the small violation conditions (90.9% vs. 92.1%, $p<0.5$), but the mean accuracy for the large violation condition (97.7%) was significantly higher than that for the other conditions ($p<0.05$).

Imaging results

Fig. 1 illustrates the three maps showing significant brain activations for each of the three sentence type conditions, relative to the resting baseline. The three conditions elicited a similar pattern of activity in a set of brain regions, including bilateral prefrontal cortex, anterior and posterior cingulate, presupplementary motor area, left anterior temporal cortex, left posterior superior/middle temporal gyrus, posterior cortical areas in parietal and occipital lobes, and also subcortical areas such as thalamus and caudate.
Fig. 2 and Table 2 show the brain areas where the BOLD signal was significantly modulated by sentence type (corrected p<0.05), as identified with the 3dANOVA2 program in AFNI. These areas were mostly in the frontal cortex, including bilateral IFG (BA 45), the left middle frontal gyrus (BA 9), bilateral precentral gyrus (BA 6), medial frontal gyrus (BA 9), and presupplementary motor area (preSMa, BA 6). There were also activations in anterior cingulate cortex (ACC, BA 32), posterior cingulate (BA 29), the left caudate, the right cuneus (BA 18), and the left lingual gyrus (BA 18).

Table 3 shows the BOLD responses (i.e., percent signal changes) computed for each sentence type in each activated region. Pair-wise comparisons indicated that for all areas except the posterior cingulate, brain activity was significantly increased in the small violation condition, relative to the large violation condition. For some areas, including bilateral IFG, the left middle frontal gyrus, medial frontal gyrus, ACC, and caudate, brain activity was comparable in the normal and small violation conditions. For other areas, including bilateral precentral gyrus, preSMa, cuneus, and lingual gyrus, brain activity was significantly greater in the normal condition than in the small violation condition.

Discussion

The present study demonstrated that participants were significantly slower in responding to the sentences with a small degree of semantic violation relative to those with a large violation. Consistent with the RT results, the accuracy for the small violation condition was 92.1%, significantly lower than the 97.7% in the large violation condition. This pattern suggests that small violations were more difficult to detect relative to large violations, and confirms that the sentences with a small violation involved more semantic integration than did those with a large violation. Furthermore, when responding to the normal sentences, the participants were significantly slower relative to the small violation sentences. However, the accuracy was comparable across the two types of sentence, indicating the absence of any speed accuracy trade-off.

Different from the present findings, many previous studies using compatible violation paradigms found worse performance in the anomalous condition than in the normal condition (e.g., Baumgaertner et al., 2002; Kiehl et al., 2002; Ni et al., 2000). There are also studies showing no difference between the normal vs. anomalous contrast or a similar pattern of the result as in the present study (e.g., in European languages: Kuperberg et al., 2003; Kuperberg et al., 2000; Kuperberg et al., 2008; Osterhout and Nicol, 1999; in Chinese: Schirmer et al., 2005; Wang et al., 2008). As mentioned in the Introduction, this complex pattern suggests that the normal vs. anomalous comparison may be subject to influence from factors unrelated to semantic processing and modulated by task demands and materials. Hence, it is not appropriate to associate performance differences with brain activity differences across the two conditions in examining semantic processing using the violation paradigm.

As we used a simultaneous presentation paradigm and the critical target words were not always placed at the end of the sentence, it may be argued that participants could detect the anomalies and responded before they reached the sentence end. Therefore, the response time differences across the three conditions could be accounted for, at least partly, by the different number of words participants actually read/processed in different conditions. However, target word location seems unlikely a critical factor. We manipulated the target word location and found this had no effect on the results (Schirmer et al., 2005). The pattern of results is the same as those of the present study: Response times were significantly slower to the congruous sentences than to the incongruous sentences.

In addition, to examine whether the manner of stimulus presentation had any effect on our results, we conducted a new behavioral experiment with the same task and materials as described in the method section except switching to a rapid serial visual presentation (RSVP) paradigm (Chen, 1986; Chen et al., 1985). For each sentence, a fixation was shown for 300 ms, followed by 200 ms blank screen, and then by the serial presentation of the words in each sentence. Each word was shown for 300 ms with a 300 ms gap between two successive words. Participants were instructed to make their responses after the full sentence had been presented. Results in this RSVP experiment with 18 participants from the same subject population were as what we found with the simultaneous presentation paradigm in the present study. The congruous condition was significantly longer than the small violation condition with a comparable accuracy (866 ± 70 ms vs. 770 ± 59 ms, p<0.05; 93.7% ± 1.2 vs. 92.7% ± 1.1, p<0.05) and the small violation condition was significantly longer and less accurate than the large violation condition.

### Table 2
Detail information for brain activations shown in Fig. 2.

<table>
<thead>
<tr>
<th>Anatomic structure</th>
<th>Coordinates (x,y,z)</th>
<th>Peak F value</th>
<th>Volume (mm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. Inferior frontal gyrus</td>
<td>45 −50 17 15</td>
<td>13.0</td>
<td>567</td>
</tr>
<tr>
<td>R. Inferior frontal gyrus</td>
<td>45 50 14 18</td>
<td>15.6</td>
<td>432</td>
</tr>
<tr>
<td>L. Middle frontal gyrus</td>
<td>9 −42 11 35</td>
<td>10.5</td>
<td>432</td>
</tr>
<tr>
<td>L. Medial frontal gyrus</td>
<td>9 −5 41 33</td>
<td>14.1</td>
<td>466</td>
</tr>
<tr>
<td>L. Precentral gyrus</td>
<td>6 −47 −2 39</td>
<td>19.2</td>
<td>756</td>
</tr>
<tr>
<td>R. Precentral gyrus</td>
<td>6 44 5 33</td>
<td>16.2</td>
<td>810</td>
</tr>
<tr>
<td>Presupplementary motor area</td>
<td>8 −2 20 45</td>
<td>18.9</td>
<td>375</td>
</tr>
<tr>
<td>Anterior cingulate cortex</td>
<td>32 −4 25 32</td>
<td>15.2</td>
<td>864</td>
</tr>
<tr>
<td>L. Caudate</td>
<td>−8 −5 9</td>
<td>14.8</td>
<td>837</td>
</tr>
<tr>
<td>R. Cuneus</td>
<td>18 11 −71 15</td>
<td>14.5</td>
<td>837</td>
</tr>
<tr>
<td>L. Lingual gyrus</td>
<td>18 −11 −77 6</td>
<td>14.6</td>
<td>459</td>
</tr>
<tr>
<td>L. Posterior cingulate</td>
<td>29 −11 −47 9</td>
<td>15.0</td>
<td>405</td>
</tr>
</tbody>
</table>

BA for Brodmann’s area. L for left and R for right. Degrees of freedom for the F values were df₁ = 2, df₂ = 30.

### Table 3
Mean percent signal changes for all three conditions in all regions of activation.

<table>
<thead>
<tr>
<th>Brain region</th>
<th>Normal</th>
<th>Small violation</th>
<th>Large violation</th>
<th>Post-hoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. Inferior frontal gyrus</td>
<td>0.38 (0.04)</td>
<td>0.38 (0.04)</td>
<td>0.30 (0.04)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>R. Inferior frontal gyrus</td>
<td>0.21 (0.03)</td>
<td>0.20 (0.03)</td>
<td>0.14 (0.02)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>L. Middle frontal gyrus</td>
<td>0.41 (0.05)</td>
<td>0.42 (0.05)</td>
<td>0.33 (0.04)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>L. Medial frontal gyrus</td>
<td>0.15 (0.03)</td>
<td>0.15 (0.04)</td>
<td>0.07 (0.03)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>Anterior cingulate cortex</td>
<td>0.26 (0.03)</td>
<td>0.26 (0.03)</td>
<td>0.19 (0.03)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>L. Caudate</td>
<td>0.31 (0.04)</td>
<td>0.27 (0.03)</td>
<td>0.20 (0.03)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>L. Precentral gyrus</td>
<td>0.34 (0.04)</td>
<td>0.30 (0.04)</td>
<td>0.25 (0.03)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>R. Precentral gyrus</td>
<td>0.20 (0.04)</td>
<td>0.16 (0.03)</td>
<td>0.11 (0.03)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>Presupplementary motor area</td>
<td>0.44 (0.05)</td>
<td>0.41 (0.04)</td>
<td>0.32 (0.04)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>R. Cuneus</td>
<td>0.30 (0.03)</td>
<td>0.26 (0.04)</td>
<td>0.22 (0.03)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>L. Lingual gyrus</td>
<td>0.31 (0.04)</td>
<td>0.26 (0.03)</td>
<td>0.22 (0.04)</td>
<td>Normal = Small&gt;Large</td>
</tr>
<tr>
<td>Posterior cingulate</td>
<td>0.31 (0.04)</td>
<td>0.22 (0.04)</td>
<td>0.22 (0.04)</td>
<td>Normal = Small=Large</td>
</tr>
</tbody>
</table>

Standard errors in brackets. *" meaning no significant differences between two conditions and **" meaning one condition being significantly larger than the other.

Marginal significance, corrected p = 0.08.
condition (770 ± 59 ms vs. 650 ± 53 ms, p < 0.01; 92.7% ± 1.1 vs. 97.2% ± 0.9, p < 0.01). This indicates that the current findings were not likely artifacts from the manner of stimulus presentation.

As manipulating the degree of violation can better control for confounding factors such as response type and task demand, the most relevant imaging result was from the small vs. large violation comparison. As indicated by the behavioral results, a small degree of semantic violation should involve more semantic integration than a large degree of violation. The brain areas that are responsible for semantic integration should, therefore, manifest an activity pattern of greater BOLD signal in the small violation condition, relative to the large violation condition. Critically, we examined whether the LIFG or the anterior temporal cortex would demonstrate such a pattern of activation, each as predicted by the MUC and the B AIS models.

The ANOVA results revealed that the LIFG was sensitive to our manipulation of sentence type and showed stronger activity for the small violation condition, relative to the large violation condition. In contrast, no anterior temporal cortical region was shown in the ANOVA, suggesting that the brain activity in this region was not modulated by sentence type and did not differentiate between the small and the large violation conditions. However, this does not imply that there was no activation in the anterior temporal cortex at all for the various conditions relative to the resting baseline. See Fig. 1 for such activations. It only means that the level of BOLD signal in this region did not differ by sentence type. This key result thus is consistent with the prediction of the MUC model, but not with the prediction of the B AIS model, suggesting that the LIFG should be the area critical for semantic integration.

Whereas no activation differences in the LIFG were found between the congruous target words and the incongruous target words that were both semantically related to the sentence context, differences were found between the incongruous target words that were related to the sentence context and the incongruous target words that were not related to the sentence context. This suggests an interesting possibility that the integration process reflected by these activation differences may rely on the knowledge of common relationships between concepts, and is not sensitive to whether the sentence makes good sense. Further research is needed to explore this possibility.

One reason that there was no main effect of sentence type in the anterior temporal cortical areas may be due to susceptibility effects as this region is known to be prone to such artifacts in BOLD fMRI data (e.g., Noppeney and Price, 2004; Devlin et al., 2000). Although there are studies which did report reliable activations in these regions without using any special protocols (e.g., Humphries et al., 2005; Willems et al., 2008), there is still a possibility that our present findings may be undermined by these artifacts. Improved acquisition, such as shorter TE or tailored RF pulses (Devlin et al., 2000) can be adopted to reduce such susceptibility artifacts in future studies. As a reference, the signal-to-noise ratio averaged over all of our participants was 12.8 ± 0.5 in the anterior temporal cortex (anterior was defined as y = −12 in Talairach space), comparable to the value for the whole brain (14.0 ± 0.43). The baseline value was 0.34 ± 0.38 for the non-brain regions. The signal quality seems to be better than some previously reported. For example, Devlin et al. (2000) found that 64% of the voxels in left anterior temporal cortex suffered a BOLD signal loss of 50% or more.

A key feature of the present study is that we followed the principle of parametric design to manipulate and compare across levels of semantic violation to avoid a direct comparison between anomalous and normal sentences which are subject to processes non-specific to semantic processing. Recently, Stanczak et al. (2007) also conducted a study along a similar line. In addition to normal sentences, they manipulated the level of plausibility and constructed anomalous and implausible sentences. While their anomalous sentences were like our large violation sentences, their implausible sentences depict rare events that can still be integrated, such as Vanessa threw the feather but did not win the competition. The implausible sentences were found to be more difficult to categorize than the anomalous sentences, with a significantly longer response time and an exceedingly low accuracy of 36%. In addition, the LIFG showed greater activity for the implausible condition, relative to the normal condition. As the so-called implausible sentences were still considered reasonably acceptable by their participants, with a mean rating of 3.9 at a 1–7 scale (1 for “highly implausible” and 7 for “highly plausible”), the manipulation in that study seems to be still within an acceptable range (i.e., the implausible sentences were mere a variation of the normal sentences). In contrast, our focus here is on the variation of the violation condition contrasting two types of non-acceptable sentences. Despite these differences, the Stanczak et al. (2007) study and the present one are consistent in demonstrating the same sensitivity of the LIFG to variations in semantic plausibility or acceptability. Combined with the assumption that such variations correspond to different levels of semantic integration, the LIFG activation observed in both studies supports the view that this area is critical for semantic integration.

It should be noted that the anterior temporal cortex may still serve some integration functions, as the B AIS model advocates, though not necessarily for processing semantics at the sentence level. For example, there are studies showing modulation of anterior temporal activity by syntactic manipulations (Humphries et al., 2006; Humphries et al., 2005). Other studies showing anterior temporal activation utilized textual stimuli, suggesting that integration may occur at the discourse level (Mazoyer et al., 1993; Xu et al., 2005). However, our data are clearly inconsistent with the idea that the anterior temporal cortex is a crucial area involving in semantic integration at a sentence level.

Although our focus was on the LIFG and the temporal cortex, the small vs. large violation contrast also revealed a number of other regions. Some of these areas, including the right inferior frontal gyrus (RIFG), the left middle frontal gyrus BA 9, medial frontal gyrus, ACC, and caudate, showed a similar activation pattern as the LIFG, with comparable brain activity across the normal and the small violation conditions. There is evidence from the literature that these areas may be closely related to linguistic processing. The RIFG area, homologous to the LIFG, may be recruited due to an increased need for integration, as suggested by some imaging studies on language (Just et al., 1996; Rodd et al., 2005), or due to searching for alternative interpretations (see a review in Stowe et al. (2005)), a function that requires more diffuse semantic processing and may involve bilateral activations (Jung-Beeman, 2005).

ACC, DLPFC BA 9, caudate and medial prefrontal cortex, in addition to being important for general cognitive processes, have also been implicated in language comprehension, such as control and evaluation (Crinion et al., 2006; Ferstl et al., 2005). Particularly, ACC and BA 9 may be related to the control process proposed in the MUC model. One alternative explanation for the BA 9 activation is that the present study used materials in the Chinese language, which is substantially different from Indo-European languages. Previous work on Chinese word processing has generally reported a similar BA 9 activation, rarely observed in English studies, and interpreted it as serving a unique role for visual processing of the Chinese script (Tan et al., 2001; Tan et al., 2000; though see Booth et al. (2006)). There are also studies on Chinese sentence comprehension (Chee et al., 1999; Mo et al., 2005; Wang et al., 2008) reporting activation in this area, suggesting that its function may not be limited to orthographic processing in Chinese.

There are other areas revealed from the small vs. large violation comparison in which brain activity was greater in the normal condition than in the small violation condition. These areas include bilateral precentral gyrus, preSMA, cuneus, and lingual gyrus. Based on the related literature, we interpret activations in these areas to be related to increased processing load in general tasks not specific to sentence-level semantic integration. The precentral gyrus, for
example, has been found to be related to task difficulty (Desai et al., 2006) and preSMA to control of actions (Aron et al., 2007; Hoshi and Tanji, 2004; Shimoda and Tanji, 2000). The visual cortical activation in lingual gyrus and the activation in cuneus can also be attributed to increase visual processing in normal sentences which were associated with a higher task load as shown from the RT data. Fu et al. (2002) reported similar visual cortical activation due to increased processing load in a lexical processing task with Chinese word stimuli. This interpretation is consistent with the present finding that the brain activity in these areas correlated positively with response time and became greater in order from the large violation, the small violation, to the normal conditions, suggesting an increase in task load across the three conditions.

Only the posterior cingulate was found to be activated significantly stronger for the normal condition than for either of the two violation conditions. This area was affected in the early course of Alzheimer’s disease that is characterized with memory deficit (Godbolt et al., 2006; Mosconi et al., 2004); and previous studies on narrative speech comprehension have interpreted its activation as an index for memory encoding and retrieval (Awan et al., 2007). It is possible that in the present study a meaningful memory representation was successfully retrieved for the normal sentences but not for the violation sentences.

It is worth mentioning that although we did not find any region showing greater activity in the two anomalous conditions than in the normal condition, as found in previous studies using the violation paradigm, such result should not be taken to indicate that there were no violation-related activities in the anomalous conditions. Given that the BOLD signal is a pooled response from more than one cognitive process, we speculate that the areas sensitive to violation may also be engaged in other processes. Therefore, violation-related activities may be mixed up with and masked by activities associated with other processes. This reinforces our point that results from direct comparison between anomalous and normal sentences should be interpreted with caution.

Conclusion

In summary, although the anomalous vs. normal comparison may identify brain regions responsible for semantic integration, we took an alternative approach to manipulate the degree of semantic violation and compared two conditions with differential semantic integration involvement in order to remove the impact of non-semantic processing. As expected, we observed increased difficulty in behavioral performance for rejecting sentences with a small semantic violation requiring more integration relative to sentences with a large violation. More importantly, we observed correspondingly stronger activity in the small violation condition, relative to the large violation condition, in a number of brain regions, among which, there was a reliable activation in the LIFG, but not in temporal cortex. Our result therefore indicates that the LIFG is crucial for integrating individual word meaning to sentence-level messages, as advocated in theories of language processing such as the MUC model (Hagoort, 2005).

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (#30670700, #30670702), the Guangdong (China) Natural Science Foundation (#805050632), Program for New Century Excellent Talents in University of China (#NCET-08-0645), the Chinese University of Hong Kong (Direct grant #20020911), and the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK4142/04H and CUHK441008). We thank QiuLin Wu, LinFa Wu, ShaoXing Chen in Medical College of Shantou University for the help in data collection, and Xuchu Weng and Zhi Yang at the Institute of Psychology of the Chinese Academy of Sciences for their help in data analysis.

References

