NEURAL BASIS OF THE SOUND-SYMBOLIC CROSSMODAL CORRESPONDENCE BETWEEN AUDITORY PSEUDOWORDS AND VISUAL SHAPES

Kelly McCormick[1,2], Simon Lacey[1,4,5], Randall Stilla[3], Lynne C. Nygaard[2] and K. Sathian[1,2,4,5,6]

Departments of Neurology[1] & Psychology[2], Winship Cancer Institute[3], Emory University, Atlanta, GA 30322, USA; Departments of Neurology[4], Neural & Behavioral Sciences[5] and Psychology[6], Milton S. Hershey Medical Center, Penn State College of Medicine, Hershey, PA 17033-0859, USA.

Short title: Neural basis of pseudoword-shape sound symbolism

Keywords: multisensory; magnitude; semantic; phonology; attention; congruency effect

Corresponding author:
K. Sathian
Department of Neurology
Milton S. Hershey Medical Center
Penn State College of Medicine
30 Hope Drive, PO Box 859, Mail Code EC037
Hershey, PA 17033-0859, USA
Tel: 717-531-1801
Fax: 717-531-0384
Email: ksathian@pennstatehealth.psu.edu

ABSTRACT

Crossmodal correspondences refer to associations between apparently unrelated stimulus features in different sensory modalities. Sound symbolism, a special class of crossmodal correspondence between the sounds of words and their meanings, has been studied by matching auditory pseudowords, e.g. ‘takete’ or ‘maluma’, with pointed or rounded visual shapes, respectively. We report here on a functional magnetic resonance imaging study in which participants were presented with audiovisual pseudoword-shape pairs that were sound symbolically congruent or incongruent. In whole-brain analyses, there were no significant neural congruency effects during attention to visual shapes, but when participants attended to the auditory pseudowords, we observed greater activity for incongruent compared to congruent (I > C) audiovisual pairs bilaterally in the intraparietal sulcus and neighboring supramarginal gyrus, and in the left middle frontal gyrus. Comparing these activations to independent functional localizers and additional region-of-interest analyses revealed no evidence for semantic mediation, and limited evidence for processes relating to multisensory integration and magnitude estimation as possible underlying mechanisms. Stronger support was found for a relationship to phonological processing and/or multisensory attention. Further, when attending to auditory pseudowords, incongruency (I > C) activation magnitudes in visual cortical foci and the precuneus were positively correlated, across participants, with preferences for visual object imagery, suggesting potential neural substrates for individual differences in sound symbolism. These findings are important for understanding a central question in neurolinguistics: the nature of sound-to-meaning mapping in the brain.

INTRODUCTION

A central question in linguistics is how language conveys meaning through the sounds of words. While the relationship between sound and meaning is often considered largely arbitrary (de Saussure, 1916/2009; Pinker, 1999; Jackendoff, 2002), there are reliable sound-meaning mappings in natural language (Blasi et al., 2016), to which language users appear sensitive (e.g., Nygaard et al., 2009). Such mappings are termed 'sound symbolic' (e.g., Knoeferle et al., 2017). The idea that words can convey meaning via auditory resemblance to their referents has a long history, an early discussion being found in Plato's Cratylus Dialog (see Ademollo, 2011). An obvious example of sound-meaning mapping is onomatopoeia, in which the sound of a word mimics the sound that the word represents (Catricalà & Guidi, 2015; Schmidtke et al., 2014), for example: 'fizzle', 'bang', or 'splash'. However, while the sound-symbolic relation in onomatopoeia is within-modal (i.e., word sounds are used to index sound-related meanings), many languages possess a distinct class of words, variously known as ideophones or mimetics, whose phonological structure is reliably mapped not only to sounds but also to sensory meanings beyond the auditory domain, i.e., crossmodally (Blasi et al., 2016; Dingemanse, 2012). For example, Japanese has a rich lexicon of mimetics including '*kirakira*' (flickering light), '*pika*' (a flash of light), and '*nurunuru*' (the tactile sensation of sliminess: Akita & Tsujimura, 2016; Kita, 1997).

Sound symbolism is often studied by matching auditorily-presented pseudowords with two-dimensional visual shapes: e.g. 'takete' or 'kiki' are matched with shapes that have pointed contours whereas 'maluma' or 'bouba' are matched with shapes bearing rounded contours (Köhler, 1929, 1947; Ramachandran & Hubbard, 2001). In this audiovisual pairing, which is classed as a type of crossmodal correspondence, the shape can be considered an abstraction of an empirically tractable object property that represents sound meaning. Crossmodal correspondences are near-universally experienced associations between ostensibly unrelated stimulus features in different sensory modalities (Spence, 2011). For example, high- and low-pitched auditory stimuli tend to be matched with visual stimuli at high and low spatial elevation, respectively (e.g., Bernstein & Edelstein, 1971; Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Lacey et al., 2016; Jamal et al., 2017), and with small- and large-sized visual stimuli, respectively (Gallace & Spence, 2006; Evans & Treisman, 2010). Interestingly, individuals with synesthesia appear to be more sensitive than non-synesthetes to sound-symbolic, but not lower-level (e.g. pitch-elevation and pitch-size) crossmodal correspondences (Lacey et al., 2016).

One possible explanation for sound symbolism is that the relationships between sound-symbolic words and their visual or semantic referents mimic audiovisual statistical regularities in the natural environment (Sidhu & Pexman, 2017). Such regularities may also underlie other crossmodal correspondences (Spence, 2011), for example that between high/low auditory pitch and high/low auditory and visual elevation (Jamal et al., 2017), since high-pitched sounds tend to emanate from high locations and low-pitched sounds from low locations (Parise et al., 2014).

Thus, sound symbolism might be connected with a more general multisensory integration process in which auditory and visual features are linked (Ković et al., 2010). If so, neural activity related to sound-symbolic processing might co-localize with activity related to multisensory integration, e.g. in the superior temporal sulcus when audiovisual synchrony (Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014) or audiovisual identity (Sestieri et al., 2006; Erickson et al., 2014) are manipulated, or in regions such as the intraparietal sulcus (IPS) when audiovisual spatial congruency is manipulated (Sestieri et al., 2006).

Sensory features can often be characterized along polar dimensions of magnitude where one end is 'more than' the other (Smith & Sera, 1992) and stimuli in different modalities might be associated by virtue of their common position along a magnitude dimension. Such crossmodal magnitude relations have been proposed as a basis for crossmodal correspondences (Lourenco & Longo, 2011; Spence, 2011; Sidhu & Pexman, 2017). In sound symbolism, magnitude might be linked to the sound structure of spoken language, in that shapes can vary along dimensions of pointedness or roundedness, and phonetic form could potentially reflect variation on related dimensions. Both visuospatial attributes of shapes (e.g., size, spatial frequency) and phonetic attributes of linguistic segments (e.g., sonority, formant frequencies) could be encoded by a domain-general magnitude system (Dehaene et al., 2003; Walsh, 2003; Lourenco & Longo, 2011) in which different attributes might become associated by virtue of occupying similar positions along magnitude dimensions. Some evidence for this comes from studies in which pseudowords with varying phonetic features were readily placed along continua of pointedness and roundedness (McCormick et al., 2015) or linearly matched with novel objects of varying size (Thompson & Estes, 2011), as well as the systematic relationship of sound-symbolic associations to phonetic features of sounds (Knoeferle et al., 2017; McCormick et al., 2015) and radial frequency patterns of shapes (Chen et al., 2016). In this case, activity related to sound symbolism might be expected in the intraparietal sulcus (IPS), an area involved in processing both numerical and non-numerical (e.g., luminance) magnitude (Sathian et al., 1999; Eger et al., 2003; Walsh, 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Sokolowski et al., 2017).

Crossmodal sound-symbolic associations exist in natural language as well as for pseudowords, for example, 'balloon' and 'spike' for rounded and pointed shapes (Sučević et al., 2015; Blasi et al., 2016). Familiar words such as these have established semantic, as well as sound-symbolic, associations. In contrast, pseudowords lack established semantic associations by definition, and thus their sound-symbolic associations may rely more on phonological processing of their sound structure. This leads to the prediction that pseudoword-shape associations could activate brain regions involved in phonological processing. However, when people attempt to match pseudowords with shapes they might also rely in part on semantic associations between the pseudowords and phonologically close real words in order to assign meaning. For example, in assigning 'maluma' to a rounded shape (Köhler, 1929, 1947), they might invoke its phonological

neighbor 'balloon'. Therefore, it is also possible that the evaluation of pseudoword-shape associations might activate regions in the left hemisphere lexical-semantic network, e.g. as identified by Fedorenko et al. (2010, 2011) using language localizers contrasting sentences with strings of pseudowords. The reverse contrast of pseudowords to sentences could be used to index phonological processing, since processing pseudoword strings does not involve the semantic or syntactic processing involved in natural sentences but only the phonological aspects of the pseudowords (Fedorenko et al., 2010).

There have been relatively few studies of the neural basis of sound symbolism, and drawing inferences about underlying mechanisms from these is problematic for several reasons. Firstly, some studies employed a variety of sound-symbolic words (Ković et al., 2010; Revill et al., 2014; Sučević et al., 2015; Lockwood et al., 2016), but there may be different mechanisms for different sound-symbolic effects (Sidhu & Pexman, 2017). Secondly, although an earlier neuroimaging study employed functional localizers for the domains of shape and motion referred to by the pseudowords that were used, this study was not designed to distinguish between potential mechanisms (Revill et al., 2014). Lastly, EEG studies (Ković et al., 2010; Sučević et al., 2015; Lockwood et al., 2016), while providing excellent temporal information, may not have sufficient anatomical resolution to define the relevant brain regions.

Furthermore, few studies have examined individual differences in crossmodal correspondences. We addressed this by administering the Object-Spatial Imagery & Verbal Questionnaire (OSIVQ: Blazhenkova & Kozhevnikov, 2009). We hypothesized that individuals with a preference for verbal processing might be more adept at assigning potential meanings to pseudowords. Additionally, object imagers might be more apt to visualize the shapes associated with pseudowords since they typically integrate multiple sources of information about an object, compared to the more schematic representations of spatial imagers.

The approach we took in the present study was to use functional magnetic resonance imaging (fMRI) to investigate cerebral cortical localization of differential activations for congruent and incongruent sound-symbolic crossmodal correspondences between auditory pseudowords and visual shapes. We chose to rely on implicit correspondences by asking participants to engage in a task that was independent of the pseudoword-shape correspondence. In order to test the relevance of the proposed mechanisms outlined above, we conducted three independent localizers in the same individuals, reflecting semantics/phonology, magnitude estimation, and multisensory integration. A potential role for one of these mechanisms would be indicated by finding cortical activations common to a particular localizer and the sound symbolic (in)congruency.

METHODS

Participants

Twenty participants took part in this study, but one was later excluded for excessive movement in the scanner (> 1.5mm), leaving a final sample of 19 (9 male, 10 female; mean age 25 years, 1 month). All participants were right-handed based on the validated subset of the Edinburgh handedness inventory (Raczkowski et al., 1974) and reported normal hearing and normal, or corrected-to-normal, vision. All participants gave informed consent and were compensated for their time. All procedures were approved by the Emory University Institutional Review Board.

Procedures

General

Five participants took part in the pseudoword-shape scans first, and then underwent three localizer scans to test potential mechanisms underlying the pseudoword-shape correspondence. The remaining participants had already undergone the localizer scans approximately four months earlier as part of a separate study and thus completed the pseudoword-shape scans after the localizers. After completing the required scan sessions, all participants performed a behavioral task to determine the strength of their crossmodal pseudoword-shape correspondence. All experiments were presented via Presentation software (Neurobehavioral Systems Inc., Albany CA), which allowed synchronization of scan acquisition with experiments and also recorded responses and response times (RTs). After the final scan session, participants also completed the OSIVQ (Blazhenkova & Kozhevnikov, 2009).

Pseudoword-shape fMRI task

We created two auditory (pseudowords) and two visual (novel two-dimensional outline shapes) stimuli, that each contrasted in perceived roundedness and pointedness based on rating studies (pseudowords: McCormick et al., 2015; visual shapes: McCormick, 2015, unpublished data). The auditory pseudowords were “lohmoh” (rounded) and “keekay” (pointed). The pseudowords were digitally recorded in a female voice using Audacity v2.0.1 (Audacity Team, 2012), with a SHURE 5115D microphone and an EMU 0202 USB external sound card, at a 44.1kHz sampling rate. The recordings were then processed in Sound Studio (Felt Tip Inc., NY), using standard tools and default settings, edited into separate files, amplitude-normalized, and down-sampled to a 22.05kHz sampling rate (standard for analyses of speech). Stimulus duration was 533ms for “keekay” and 600ms for “lohmoh”. The visual stimuli were gray outline shapes on a black background (Figure 1a), each subtending approximately 1° of visual angle and presented at the center of the screen for 500ms. The selected stimuli lay near the ends of independent rounded and pointed dimensions in each modality based on empirical ratings (McCormick et al., 2015), i.e., the chosen rounded pseudowords and shapes were rated towards the high end of the rounded dimension and the low end of the pointed dimension, and vice versa for the chosen pointed stimuli. These stimuli were presented concurrently in audiovisual pairs (Figure 1a) that were either congruent (“keekay”/pointed shape or “lohmoh”/rounded shape) or incongruent (“keekay”/rounded shape or “lohmoh”/pointed shape) with respect to the crossmodal pseudoword-shape (sound-symbolic) correspondence. A mirror angled over the head coil

enabled participants to see the visual stimuli projected onto a screen placed in the rear magnet aperture. Auditory stimuli were presented via scanner-compatible headphones.

There were four runs, each consisting of 8 task blocks (4 congruent and 4 incongruent, each block containing 10 word-shape stimuli with a 3s interval between the onset of successive stimuli), each lasting 30s and alternating with 9 rest blocks, each lasting 16s; total run duration was 384s. Pseudoword-shape pairs were pseudorandomly interleaved (no more than three trials in a row of the same pairing, no more than two blocks in a row of the same condition). In two of the runs, participants attended to auditory stimuli; in the other two runs, they attended to visual stimuli; the order of attended modality was counterbalanced across participants, who performed a 2AFC task in the attended modality. Participants pressed one of two buttons on a hand-held response box when they heard either "keekay" or "lohmoh" (attend auditory condition) or saw either the rounded or pointed shape (attend visual condition). The right index and middle fingers were used to indicate responses, counterbalanced between subjects across modalities and across rounded and pointed stimuli.

Localizer tasks

The order of the localizer tasks was fixed, progressing from the one perceived as most difficult to the easiest: participants completed the magnitude estimation localizer first, then the temporal synchrony localizer, and finally the semantic localizer. Each localizer comprised two runs with a fixed stimulus order; the order of runs was counterbalanced across participants. The localizers were completed in a single session for most participants. We chose localizers that broadly reflected the three potential mechanisms underlying crossmodal correspondences proposed by Spence (2011): structural, e.g., driven by intensity or magnitude; statistical, i.e., features that regularly co-occur in the world and thus might lend themselves to multisensory integration; and semantic, i.e. driven by linguistic factors. Three of the five potential mechanisms proposed by Sidhu & Pexman (2017) are also relevant here: their Mechanism 1 relates to statistical co-occurrence, Mechanism 3 to neural codes for properties such as magnitude, and Mechanism 5 to semantic or language patterns (Mechanisms 2 and 4 relate to shared properties of phonemes and other stimuli and species-general associations respectively and are beyond the scope of the present study). A number of different localizers could be designed to test each mechanism; thus it would be difficult to test all possibilities in a single study. We made choices that seemed reasonable, with the thinking that the results of the present study should help to refine such choices in future work (see Discussion).

Semantic localizer: In this task, adapted from Fedorenko et al. (2010), we contrasted complete semantically and syntactically intact sentences with sequences of pseudowords (examples are provided in Figure 1b) to identify brain regions processing word- and sentence-level meaning (Fedorenko et al., 2010, 2011; Bedny et al., 2011). The complete sentences and pseudoword sequences each contained 12 words, each word/pseudoword being presented visually for 450ms

for a total stimulus duration of 5.4s, followed by a 600ms inter-stimulus interval (ISI) during which participants were visually prompted to press a button. There were two runs, each consisting of 16 task blocks (8 of each type, each block containing 3 stimuli) of 18s duration and alternating with 17 rest blocks of 12s duration; total run duration was 492s. We expected the contrast of complete sentences > pseudowords to reveal regions mediating both semantic and syntactic processing (Fedorenko et al., 2010, 2011; Bedny et al., 2011), i.e., a largely left hemisphere network comprising the inferior frontal gyrus (IFG), angular gyrus (AG), and extensive sectors of the temporal lobe including the superior temporal sulcus (STS). We also reasoned that the reverse contrast of pseudowords to sentences would identify regions involved in phonological processing (Fedorenko et al., 2010), given the greater role of such processing in reading pseudowords compared to sentences; however, note that Fedorenko et al. (2010) did not explicitly describe activations found on this contrast.

Multisensory integration: Among a number of possible localizers that could be used to test multisensory integration, we chose one that is sensitive to the synchrony of auditory and visual stimuli, as used in many studies of audiovisual integration (e.g., Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014). The auditory stimulus was an 810Hz tone of 800ms duration with a 20ms on/off ramp. The visual stimulus was a gray circle (RGB values 240, 240, 240) subtending approximately 1° of visual angle and presented centrally for 800ms. In synchronous trials, auditory and visual stimuli were presented simultaneously for 800ms followed by a 3200ms ISI, while asynchronous trials contained stimuli offset by an intervening blank interval of 200ms followed by a 2200ms ISI (Figure 1c). Half the asynchronous trials presented the auditory stimulus first and half the visual stimulus first. There were two runs, each consisting of 12 active 16s-blocks (6 of each type, each block containing 4 trials) and alternating with 13 rest blocks each lasting 14s; total run duration was 374s. Participants had to press a button whenever an oddball stimulus (e.g., Crottaz-Herbette & Menon, 2006), either a square or a burst of white noise, occurred; two oddballs of each type occurred in each run, one in a synchronous block and one in an asynchronous block. The contrast between synchronous and asynchronous trials was used to identify brain regions sensitive to audiovisual synchrony; we anticipated that this contrast would activate superior temporal cortex (Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; Noesselt et al., 2012; Erickson et al., 2014).

Magnitude estimation: In order to identify brain regions sensitive to magnitude, we used a modified form of the estimation task from Lourenco et al. (2012). In each trial of this task, participants were asked to estimate whether there are more black or white elements in a visual array of small rectangles. For a control task, we modified these arrays so that one item was a triangle and participants indicated whether the triangle was black or white, thus the response – black or white – was the same for both the magnitude and the control tasks (Figure 1d). There were two runs, each containing 12 active 16s-blocks (6 of each type, each block containing 4

stimuli displayed for 1s with a 3s ISI) and alternating with 13 rest blocks each lasting 14s; total run duration was 374s. The contrast of magnitude estimation > control was used to identify those regions sensitive to magnitude, which we expected would be primarily in posterior parietal cortex, particularly the intraparietal sulcus (IPS) (Sathian et al., 1999; Eger et al., 2003; Walsh, 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Lourenco & Longo, 2011; Sokolowski et al., 2017).

Post-scan behavioral testing

As a final step, we tested whether participants reliably demonstrated the cross-modal pseudoword-shape correspondence using the implicit association test (IAT: Greenwald et al., 1998; Parise & Spence, 2012; Lacey et al., 2016). Originally devised as a test of social attitudes (Greenwald et al., 1998), the IAT has been successfully used to test the very different associations involved in crossmodal correspondences, including sound-symbolic ones (Peiffer-Smadja, 2010, unpublished thesis, Université Paris Descartes; Parise & Spence, 2012; Lacey et al., 2016). The underlying principle is the same: response times (RTs) are faster if the stimuli assigned to a particular response key are congruent and slower if they are incongruent (Greenwald et al., 1998; Parise & Spence, 2012). The advantage of the IAT for testing crossmodal correspondences is that presenting each stimulus in isolation eliminates the confound of selective attention effects potentially causing slower RTs for incongruent pairings (Parise & Spence, 2012).

The auditory and visual stimuli were the same as for the imaging experiment except that rounded and pointed pseudowords and shapes were presented in isolation (note that the absolute size of the visual shapes was altered so that they subtended 1° of visual angle in both the IAT and fMRI experimental set-ups). The IAT was presented via Presentation software (Neurobehavioral Systems Inc., Albany CA) which also recorded RTs. Participants were instructed to associate pairs of stimuli with one of two response keys (the ‘left’ and ‘right’ arrow keys on a normal US ‘QWERTY’ keyboard). The pairs always consisted of one auditory and one visual stimulus and, in separate blocks of trials, were either congruent or incongruent. A trial consisted either of an auditory presentation (either “keekay” or “lohmoh”) or a visual presentation (either the rounded or the pointed shape) and participants were asked to respond as quickly as possible.

There were four runs, each comprising 96 trials divided into two blocks of 48 (totaling 384 trials across runs). The response key associations were congruent in one block and incongruent in the other; there were different associations between keys and responses in each block (see below) and block order was counterbalanced across runs, thus avoiding order effects. Each block was preceded by an instruction screen describing the response key associations to be used, and by 12 practice trials (not included in the analysis) with on-screen feedback as to accuracy (for the practice trials only). For two runs, the congruent pairs were “keekay”/pointed shape (to be associated with the left arrow key) and “lohmoh”/rounded shape (associated with the right arrow

key) while the incongruent pairs were “keekay”/rounded shape (left arrow key) and “lohmoh”/pointed shape (right arrow key). On the other two runs, the left/right key associations were reversed. Half the trials were auditory (either “keekay” or “lohmoh”) and half were visual (either the rounded or the pointed shape), split equally into congruent and incongruent blocks and occurring in pseudorandom order within each block. Trials consisted of a blank 1000ms followed by either a visual or an auditory stimulus and were terminated either by the participant pressing a response key or automatically 3500ms after stimulus onset if no response was made. Unless terminated by an earlier response, visual stimulus duration was 1000ms and auditory stimulus duration was the length of the pseudoword (533ms for “keekay” and 600ms for “lohmoh”); RTs were measured from stimulus onset. The length of each active block thus varied between participants but was a maximum of 330s.

Image acquisition

MR scans were performed on a 3 Tesla Siemens Trio TIM whole body scanner (Siemens Medical Solutions, Malvern, PA), using a 12-channel head coil. T2*-weighted functional images were acquired using a single-shot, gradient-recalled, echoplanar imaging (EPI) sequence for BOLD contrast. For all functional scans, 34 axial slices of 3.1mm thickness were acquired using the following parameters: repetition time (TR) 2000ms, echo time (TE) 30ms, field of view (FOV) 200mm, flip angle (FA) 90°, in-plane resolution 3.125×3.125mm, and in-plane matrix 64×64. High-resolution 3D anatomic images were acquired using an MPRAGE sequence (TR 2300ms, TE 3.9ms, inversion time 1100ms, FA 8°) comprising 176 sagittal slices of 1mm thickness (FOV 256mm, in-plane resolution 1×1mm, in-plane matrix 256×256). Once magnetic stabilization was achieved in each run, the scanner triggered the computer running Presentation software so that the sequence of experimental trials was synchronized with scan acquisition.

Image processing and analysis

Image processing and analysis was performed using BrainVoyager QX v2.8.4 (Brain Innovation, Maastricht, Netherlands). In individual analyses, each participant’s functional runs were real-time motion-corrected utilizing Siemens 3D-PACE (prospective acquisition motion correction). Functional images were preprocessed employing cubic spline interpolation for slice scan time correction, trilinear-sinc interpolation for intra-session alignment of functional volumes, and high-pass temporal filtering to 2 cycles per run to remove slow drifts in the data without affecting task-related effects. Anatomic 3D images were processed, co-registered with the functional data, and transformed into Talairach space (Talairach & Tournoux, 1988). Talairach-normalized anatomic data sets from multiple scan sessions were averaged for each individual, to minimize noise and maximize spatial resolution.

For group analyses, the transformed data were spatially smoothed with an isotropic Gaussian kernel (full-width half-maximum 4mm). The 4mm filter is within the 3-6mm range recommended to reduce the possibility of blurring together activations that are in fact

anatomically and/or functionally distinct (White et al., 2001), and the ratio of the smoothing kernel to the spatial resolution of the functional images (1.33) matches that of studies in which larger smoothing kernels were used (Mikl et al., 2008). Runs were percent signal change normalized (i.e., the mean signal value for each voxel's time course was transformed to a value of 100, so that the individual values fluctuated around that mean as percent signal deviations).

For group activation display, we created a group average brain. We first selected a representative (target) Talairach-normalized brain from the 19-participant group. We then individually aligned the 18 remaining participants' Talairach-normalized brains to this target (co-registration to match the gyral and sulcal pattern, followed by sinc interpolation). These 18 aligned brains were then averaged. This 18-subject average brain was then averaged with the target brain, creating a single Talairach template, with 1mm isotropic resolution, which was used to display the activations for the 19-subject group. This group average brain was displayed using the real-time volume rendering option in BrainVoyager QX. For statistical analysis, the 19-subject Talairach template was manually segmented in order to create a group average cortical ‘mask’ file with 3mm spatial resolution, equivalent to the spatial resolution of the functional data files.

Statistical analyses of group data used general linear models (GLMs) treating participant as a random factor (so that the degrees of freedom equal n-1, i.e. 18), followed by pairwise contrasts. This analysis allows generalization to untested individuals. Correction for multiple comparisons within a cortical mask (corrected $p < .05$) was achieved by imposing a threshold for the volume of clusters comprising contiguous voxels that passed a voxel-wise threshold of $p < 0.001$, using a 3D extension (implemented in BrainVoyager QX) of the 2D Monte Carlo simulation procedure described by Forman et al. (1995). This stringent voxel-wise threshold is recommended to avoid potential problems of false positives and also permits more accurate spatial localization of activation clusters than is possible with more lenient thresholds (Woo et al., 2014; Eklund et al., 2016). Activations were localized with respect to 3D cortical anatomy with the help of an MRI atlas (Duvernoy, 1999).

In reporting activations, we do not provide ‘hotspot’ coordinates (i.e. the voxel with the largest t-value) because the statistical significance of specific voxels is not tested against other voxels within the activation (Woo et al., 2014). Instead, we provide the ‘center of gravity’ (CoG) coordinates since these orient the reader to the anatomical location but are statistically neutral. Where an activation spans several anatomical locations, we provide an extended description (see Table 1). Likewise, in order to compare the present results to previous studies, we have visually inspected activations reported in those papers rather than compute the Euclidean distance between coordinates since this would have involved relying on ‘hotspot’ coordinates in most cases.

RESULTS

Behavioral

In-scanner tasks

In the semantic localizer, participants were equally accurate in responding to the visual cue at the end of each sentence (mean ± sem: 98.4±1.0%) or pseudoword string (97.7±1.1%; t_{18} = .9, p =.4). Because of the low number of oddball trials (four for each participant) in the multisensory localizer, we conducted a non-parametric Wilcoxon test, which showed no significant difference between detection of synchronous (90.1±3.9%) and asynchronous (85.5±5.5%) oddballs (Z = -1.4, p = .2). For the magnitude localizer, there was no significant difference in accuracy between the magnitude estimation (92.2±2.0%) and control (96.4±1.3%) tasks (t_{18} = -1.8, p = .1) although RTs were significantly faster for the control task (900±45ms) compared to the magnitude task (991±53ms; t_{18} = 3.4, p < .01).

Pseudoword-shape task: Two-way (congruency, modality) repeated-measures ANOVA (RM-ANOVA) showed that mean (±sem) accuracy was not significantly different between the 'attend auditory' (97.4±0.7%) and 'attend visual' conditions (97.6±0.6%: $F_{1, 18}$ = .05, p = .8), nor between congruent (97.7±0.5%) and incongruent (97.3±0.6%) trials ($F_{1, 18}$ = .9, p = .3). There was a significant interaction between the attended modality and congruency ($F_{1,18}$ = 9.6, p = .006); this was due to greater accuracy for congruent compared to incongruent trials in the 'attend auditory' condition (98.2% vs 96.5%; t_{18} = 2.6, p = .018) but not the 'attend visual' condition (97.1% vs 98%; t_{18} = -1.9 p = .06: note that we are only concerned with these two comparisons, thus Bonferroni-corrected alpha = .025). We also performed non-parametric tests on the accuracy data since the 'ceiling' effects in all conditions indicate that the data were likely not normally distributed: these confirmed the result of the RM-ANOVA.

Analysis of RTs excluded trials for which there was no response (.6% of all trials) or an incorrect response (2.5% of responses), and further excluded trials for which the RT was more than ± 2.5 standard deviations from the individual participant mean (2.6% of correct response trials). RM-ANOVA showed that RTs were faster for the 'attend visual' (474±21ms) compared to the 'attend auditory' condition (527±23ms: $F_{1,18}$ = 21.3, p < .001) and for congruent (489±21ms) compared to incongruent trials (513±22ms: $F_{1,18}$ = 18.7, p < .001). There was a significant interaction between the attended modality and congruency ($F_{1,18}$ = 9.2, p < .007) in which both auditory (545ms) and visual (480ms) incongruent RTs were slower than congruent RTs in the same modality (auditory 509ms, t_{18} = -4.0, p = .001; visual 469ms, t_{18}, p = .009) but the absolute difference was greater for auditory than visual RTs (36ms vs 11ms).

Note that, despite the highly repetitive nature of the pseudoword and shape stimuli, it is unlikely that participants stopped paying attention to them given the high accuracy rates (> 97% in all conditions) and the fact that responses were made on 99.4% of trials.

Post-scan pseudoword-shape IAT

An RM-ANOVA with factors of modality (auditory pseudowords, visual shapes) and response key association (congruent, incongruent) showed that accuracy was higher for the auditory pseudowords (95.8±0.8%) compared to the visual shapes (91.9±0.8%: $F_{1,18} = 31.8$, $p < .001$) and when response key associations were congruent (95.3±0.7%) compared to incongruent (92.4±1.2%: $F_{1,18} = 5.7$, $p = .03$). The modality x response key association interaction was not significant ($F_{1,18} < .01$, $p = .9$).

Analysis of RTs excluded trials for which there was no response, or which failed to log (.7% of all trials), or incorrect responses (6.2% of responses), and further excluded trials for which the RT was more than ± 2.5 standard deviations from the individual participant mean (2.7% of correct response trials). RM-ANOVA showed that RTs were faster for the visual (606±19ms) compared to the auditory (702±21ms) stimuli ($F_{1,18} = 31.15$, $p < .001$) and when the response key associations were congruent (580±18ms) compared to incongruent (728±22ms: $F_{1,18} = 64.5$, $p < .001$). The modality x response key association interaction was not significant ($F_{1,18} = .5$, $p = .5$).

Imaging

Localizer tasks

Semantic

As expected, the contrast of complete sentences > pseudowords within the group average cortical mask described in the Methods section (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 9 voxels) revealed large activations bilaterally along the STS, extending into parts of the superior temporal gyrus (STG); this activation extended more posteriorly on the left than the right. Additional activations were noted on the left precentrally and in the middle occipital gyrus (Table 1a; Figure 2). Thus, the semantic localizer broadly replicated the previously reported language network (Fedorenko et al., 2010, 2011).

The reverse contrast of pseudowords > complete sentences within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 9 voxels) revealed large, bilateral frontoparietal activations: in and around the superior frontal sulcus (SFS) extending all the way into the inferior frontal gyrus (IFG) on the right, and in the angular gyrus (AG) extending into the supramarginal gyrus (SMG) on the left and the intraparietal sulcus (IPS) on the right. Smaller activations were also found along the medial surface of the left hemisphere in frontal and posterior cingulate cortex and in the parieto-occipital fissure (POF) (Table 1b; Figure 2).

As noted earlier, the contrast of pseudowords > sentences may reflect the greater role of phonological processing in reading pseudowords compared to normal sentences (Fedorenko et al., 2010). Consistent with a phonological basis for activation in the present study, the left IFG activation on this contrast is close to an IFG focus showing greater activity for, as here, visually presented pseudowords relative to concrete words (Binder et al., 2005). Additionally, the left SMG activation on this contrast is at a site reported to be involved in phonological processing for

visually presented words (Price et al., 1997; Wilson et al., 2011). However, reading pseudowords is also more effortful than reading complete sentences, perhaps due to mapping unfamiliar orthographic representations to phonological representations. Therefore, another possibility is that some or all of the activations on this contrast might reflect this additional effort, particularly since many of the activations (principally those in the frontal cortex bilaterally) were also found in regions identified as part of the frontoparietal domain-general, 'multiple demand' system (Duncan, 2013; Fedorenko et al., 2013).

Multisensory synchrony

The contrast of synchronous > asynchronous within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 8 voxels) revealed bilateral activations in the anterior calcarine sulcus extending through the POF to the cuneus; in the left hemisphere, this activation further encompassed foci in the lingual and posterior cingulate gyri (Table 1c; Figure 3). The reverse contrast, asynchronous > synchronous, resulted in two right hemisphere activations, one in inferior parietal cortex extending across the SMG and AG and into the IPS, and one in the IFG (Table 1d; Figure 3). While some previous studies have shown greater activation for synchronous compared to asynchronous audiovisual stimuli, others have shown the reverse, and in some cases preference for synchrony and asynchrony occurred at foci in proximity to each other (e.g. Stevenson et al., 2010). The regions most consistently reported as sensitive to synchrony in previous studies were in the superior temporal sulcus and/or gyrus – we did not observe activations in these regions on the multisensory synchrony localizer here, perhaps reflecting differences in stimuli and tasks (see Discussion). However, the right IFG area that was more active for asynchronous than synchronous stimuli in the present study was close to a region identified on the meta-analysis of Erickson et al. (2014) as exhibiting a preference for audiovisual stimuli characterized by either content incongruency or asynchrony.

Magnitude estimation

The contrast of magnitude estimation > control within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 7 voxels) showed exclusively right hemisphere activity in the SMG, the SPG extending into the IPS, and the middle occipital gyrus (MOG) extending through the intra-occipital sulcus (IOS) to the superior occipital gyrus (SOG: Table 1e; Figure 4). These loci are consistent with activations reported in previous studies of magnitude processing (Eger et al., 2003; Pinel et al., 2004; Piazza et al., 2004, 2007) and a recent meta-analysis (Sokolowski et al., 2017). The right MOG region is close to foci previously implicated in subitizing (Sathian et al., 1999) or that showed adaptation to magnitude (Piazza et al., 2007) while the right SPG-IPS region is close to a region involved in counting visual objects (Sathian et al., 1999).

<u>Pseudoword-shape task</u>

To test for regions involved in processing the sound-symbolic word-shape correspondence, we tested for voxels showing differential activation for congruent and incongruent trials (C > I or I > C), i.e., congruency or incongruency effects, respectively. Within the group average cortical mask, at a voxel-wise threshold of $p < .001$, there were no activations that survived correction for multiple comparisons globally, irrespective of the attended modality, or in the 'attend visual' condition, either for the C > I contrast or its reverse, the I > C contrast.

However, in the 'attend auditory' condition, while there was similarly no effect favoring the congruent over the incongruent condition, several regions showed greater activation for incongruent, compared to congruent, trials (the I > C contrast) within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 5 voxels). These regions included bilateral foci in the anterior IPS extending on the left into the mid-IPS and the SMG, and activations in the right SMG extending into the postcentral sulcus, in the left SPG and the left MFG (Table 2; Figure 5). Representative time-course curves for the I > C contrast in these regions are displayed in Figure 6A. These show greater activation for incongruent compared to congruent trials in all regions except for the left SPG region in which there was differential *de*activation. Differential activation and deactivation are likely to reflect different underlying mechanisms and, since the left SPG focus also did not show an overlap with any of the localizers, we do not consider it further. This neural (in)congruency effect was associated with a behavioral congruency effect in the 'attend auditory' condition in which responses to congruent trials were faster and more accurate than those to incongruent trials; this is consistent with the processing of incongruent trials being more effortful. The absence of a neural (in)congruency effect in the 'attend visual' condition is consonant with the absence of a congruency effect for accuracy data in this condition; although there was a congruency effect for RTs, this effect was significantly smaller in the 'attend visual' condition compared to the 'attend auditory' condition (see above).

Congruency effect overlap with localizers

Our approach to distinguishing between competing potential mechanisms was to look for overlap between areas showing pseudoword-shape (in)congruency effects and areas revealed by the semantic, magnitude, and multisensory localizers. Note that we compared pseudoword-shape and localizer maps when both had a voxel-wise threshold of $p < .001$. Overlaps at this strict threshold, avoiding potential false positives and allowing more accurate spatial localization than at more liberal thresholds (Woo et al., 2014; Eklund et al., 2016), would provide support for candidate mechanisms. However, the presence of an overlap does not guarantee that the same process underlies both tasks at that locus nor that the same neuronal populations are activated; moreover, note that absence of overlaps would not allow such mechanisms to be definitively ruled *out*. The evidence of overlaps should be regarded as indicative and supportive of further research into the functional roles of those loci.

None of the regions showing sensitivity to the pseudoword-shape correspondence overlapped or were contiguous with (i.e. shared a common edge or vertex) any area revealed by the semantic localizer contrasts of sentences > pseudowords or the multisensory contrast of synchronous > asynchronous (Figure 7). Instead, regions showing pseudoword-shape incongruency effects showed overlaps and contiguities with the control conditions from the semantic and multisensory localizers, i.e. when the primary contrasts for these conditions were reversed (see notes to Table 2; Figure 7). The contrast of pseudowords > complete sentences from the semantic localizer revealed overlap with incongruency effects in the left MFG and SMG. Comparing the BOLD signal time-courses for these regions (Figure 6A & B) suggests that the left SMG region may be the more relevant of the two since it showed greater activation for both incongruent sound-symbolic and pseudoword trials, relative to the corresponding comparison conditions. By contrast, the left MFG showed different response patterns in the sound-symbolic and semantic localizer tasks: while there was greater activation for incongruent than congruent trials in the former task, it barely responded to pseudowords while deactivating to the sentence trials in the semantic localizer.

Although the pseudoword condition only involves processing the phonological form of the pseudowords and not the semantic or syntactic aspects of natural language (Fedorenko et al., 2010), the pseudowords > sentences contrast may have limitations as a formal test for phonological processing (see Discussion). Therefore, we also conducted a region-of-interest (ROI) analysis in which we created bilateral SMG ROIs, chosen because they were shown to be functionally involved in phonological processing using transcranial magnetic stimulation (TMS; Hartwigsen et al., 2010). The ROIs consisted of cubes of 15mm side, centered on coordinates from Hartwigsen et al. (2010; Talairach coordinates -45, -37, 42; 45, -36, 42 – MNI coordinates were transformed into Talairach space using the online tool provided by Lacadie et al., 2008). Although the SMG was part of the inferior parietal cluster of sound-symbolic incongruency-related activations in both hemispheres, the ROI approach allows a more specific test of the relationship to phonology. The contrast of incongruent > congruent in the 'attend auditory' condition was significant in both left (t_{18} = 3.6, p = .002) and right (t_{18} = 3.9, p = .001) SMG ROIs. For completeness, there were no significant effects in either ROI in the 'attend visual' condition (left: t_{18} = 1.1, p = .3; right: t_{18} = .8, p = .5), matching the absence of activations within the cortical mask.

In the multisensory localizer, the contrast of asynchronous > synchronous showed overlaps with the right aIPS and SMG incongruency regions (see notes to Table 2; Figure 7). BOLD signal time-courses for these regions showed that both exhibited similar response patterns for the sound-symbolic and localizer tasks: greater activation for incongruent and asynchronous trials, respectively, relative to the corresponding comparison conditions (Figure 6A & C). Since the multisensory localizer did not reveal the STS activity that was expected on the basis of previous studies (Beauchamp, 2005a,b; van Atteveldt et al., 2007; Stevenson et al., 2010; Marchant et al.,

2012; Noesselt et al., 2012; Erickson et al., 2014), we also carried out an ROI analysis in bilateral STS ROIs, chosen because they were sensitive to audiovisual integration of non-speech stimuli (given that the current experimental stimuli were also non-speech, albeit of a different type), and were also behaviorally relevant in that activation profiles predicted task performance (Werner & Noppeney, 2010). These ROIs, which were also used in a prior study from our group on the pitch-elevation crossmodal correspondence (McCormick et al., 2018), comprised cubes of 15mm side, centered on coordinates from Werner & Noppeney (2010; Talairach coordinates -57, -41, 14; 52, -33, 9 – MNI coordinates were transformed into Talairach space as above). However, the contrast of incongruent > congruent was not significant in these ROIs, either in the 'attend auditory' condition (left: $t_{18} = .3$, $p = .8$; right: $t_{18} = .4$, $p = .7$) or the 'attend visual' condition (left: $t_{18} = 1.6$, $p = .1$; right: $t_{18} = 1.2$, $p = .2$).

Finally, in the magnitude localizer, the main magnitude estimation > control contrast overlapped with the right SMG incongruency region and was also contiguous with the right aIPS incongruency region (see notes to Table 2; Figure 7). The BOLD signal time-course for the right SMG region showed that this exhibited similar response patterns for the sound-symbolic and localizer tasks: greater activation for incongruent and magnitude estimation trials, respectively, relative to the corresponding comparison conditions (Figure 6A & D).

Correlation analyses

Finally, we tested whether pseudoword-shape activation magnitudes correlated, across participants, with the magnitude of individual congruency effects derived from in-scanner RTs. The congruency effect computation used the formula $(RT_i - RT_c) / (RT_i + RT_c)$, where RT_i and RT_c represent RT for incongruent and congruent trials, respectively. We also tested for correlations with scores on the verbal sub-scale of the OSIVQ (Blazhenkova & Kozhevnikov, 2009) and with the difference between scores on the object and spatial imagery sub-scales (OSdiff score: the spatial score is subtracted from the object score to give a single scale on which negative scores indicate a preference for spatial imagery and positive scores a preference for object imagery). To avoid circularity, we conducted these correlation tests in the whole brain, i.e. independently of the activations, and in a similarly stringent manner, by setting a strict voxel-wise threshold of $p < 0.001$ within the cortical mask before applying cluster correction (corrected $p < 0.05$).

In the 'attend visual' condition, activation magnitudes for the C > I contrast were negatively correlated with the in-scanner RT congruency magnitudes at a focus in the left SMG (Table 3a; Figure 8); this focus overlapped with the left SMG focus in the semantic control task. Additionally, a single voxel in the right SFS, contiguous with right SFS activation on the semantic control task (Table 3b; Figure 8), had a C > I activation magnitude that was positively correlated with verbal scores from the OSIVQ.

In the 'attend auditory' condition, several foci showed strong positive correlations between activation magnitudes for the I > C contrast and in-scanner RT congruency magnitudes (Table 3c; Figure 8), overlapping with the 'attend auditory' incongruency region in the left aIPS/SMG and close to the left mid-IPS incongruency region. Among the localizer regions, correlation foci were contiguous with semantic control activations in the right precuneus and the left AG, IPS, and SMG, and close to the semantic control activation in the right IFS/IFG. A single correlation focus was close to the right SPG activation from the magnitude localizer. There were also foci at which incongruency (I > C) magnitudes were positively correlated with preference for visual object imagery (Table 3d; Figure 8). These correlated regions were primarily in the left precuneus, and visual cortical areas: the IOS, the posterior calcarine sulcus extending into the lingual gyrus, and the right MOG.

DISCUSSION

To our knowledge, the present study is the first to investigate in a principled way the neural mechanisms *by* which sound-symbolic associations are processed, as opposed to merely the neural locus *at* which they are processed. Previous studies have employed a range of sound-symbolic words and pseudowords (Ković et al., 2010; Revill et al., 2014; Sučević et al., 2015; Lockwood et al., 2016) which may rely on various mechanisms depending on the particular sound-symbolic relation (Sidhu & Pexman, 2017). Here, we concentrated on a specific kind of association, that between auditory pseudowords and visual shapes, investigating the effect of manipulating the congruency of this sound-symbolic correspondence. We compared the resulting neocortical activations to the results of functional localizers designed to reflect potential underlying mechanisms. This was in contrast to earlier studies which simply relied on reverse inference in the absence of localizer scans (Peiffer-Smadja, 2010, unpublished thesis, Université Paris Descartes) or used localizers reflecting the perceptual domains to which the pseudowords were intended to refer, rather than specifically addressing the underlying mechanisms (Revill et al., 2014). Finally, we employed fMRI which offers greater anatomical resolution than the EEG paradigms of some prior studies (e.g.,Ković et al., 2010; Sučević et al., 2015; Lockwood et al., 2016). As noted above, our findings are subject to the caveat that overlaps between functional localizers and the pseudoword-shape task activations serve as a spur to further research rather than guaranteeing that the mechanism has been conclusively identified.

Relationship of incongruency effects to localizers

Semantic and phonological processes

The language regions identified on the semantic localizer showed no overlap at all with the observed activations due to incongruent, compared to congruent, pairings of auditory pseudowords and visual shapes. Thus, of the *a priori* explanations considered in the Introduction, the possibility of semantic mediation, i.e. correspondence between meanings that might be implicit in the pseudowords (by reference to relevant phonologically neighboring words) and the shapes that are associated, can be essentially discounted.

However, the incongruency-related region in the left SMG and MFG overlapped with sites more active for pseudowords than complete sentences in the semantic localizer. In addition, foci demonstrating correlations of neural incongruency effects with in-scanner RT incongruency effects during the 'attend auditory' condition, in the right precuneus and right inferior frontal cortex, were adjacent to or near areas with a preference for pseudowords over sentences.

The premise of sound symbolism is that the sound structure of words mimics or relates to some aspect of what they represent, either explicitly and within-modally as in onomatopoeia (Catricalà & Guidi, 2015; Schmidtke et al., 2014) or by analogy as for ideophones or mimetics (Akita & Tsujimura, 2016; Kita, 1997) – for example, the repetitive sound of the Japanese mimetic term 'kirakira' may be considered analogous to flickering light. It is thus interesting that we found a left-lateralized regions in the SMG and MFG that were common to both the pseudoword-shape incongruency effect and the pseudoword condition of the semantic localizer, even though the former were presented auditorily and the latter were presented visually. Given that reading pseudowords depends almost entirely on phonological processing (Fedorenko et al., 2010) whereas reading sentences also involves syntactic and semantic processing, this suggests that phonological processes may underlie this overlap. Both the left and right SMG are involved in phonological processing (Hartwigsen et al., 2010; Oberhuber et al., 2016), and the left SMG focus has previously been implicated in phonological processing for visually presented words (Price et al., 1997; Wilson et al., 2011). In order to test the phonology hypothesis further, we conducted an ROI analysis of bilateral SMG foci derived from an earlier study (Hartwigsen et al., 2010), i.e., independently of our findings. This ROI analysis showed incongruency effects in bilateral SMG in the 'attend auditory', but not visual, condition, further supporting a role for phonological processing in the pseudoword-shape correspondence.

There are many aspects to phonology, e.g., syllable structure, stress, and prosody, and while the pseudowords > sentences contrast may reflect phonological processing, it likely involves multiple elements that need to be disentangled. Although several studies suggest that consonants contribute more than vowels to associations with pointed and rounded shapes (e.g., Nielsen & Rendall, 2011; Ozturk et al., 2013; Fort et al., 2015; but see Styles & Gawne, 2017) and that pointed/rounded shapes are associated with plosive/sonorant consonants (Monaghan et al., 2012; Fort et al., 2015), further work is required to investigate the relative contributions of different phonological and/or phonetic features to particular sound-symbolic associations (e.g. Knoeferle et al., 2017; McCormick et al., 2015). Finally, it should be noted that reading pseudowords involves mapping unfamiliar orthographic representations to phonological representations and may thus require more effort than reading complete sentences; therefore, the overlaps with the activations related to sound symbolism may reflect involvement of the domain-general multiple demand system (Duncan, 2013; Fedorenko et al., 2013), rather than, or in addition to, phonological processes.

Multisensory and magnitude estimation processes

As described in the Results, the multisensory localizer did not show activity in superior temporal cortex, the region most commonly activated in prior studies employing audiovisual (a)synchrony to demonstrate multisensory integration. Most previous studies employed audiovisual speech (van Atteveldt et al., 2007; Stevenson et al., 2010; Noesselt et al., 2012; Erickson et al., 2014) or combinations of familiar environmental sounds and images (e.g., Hein et al., 2007; Noppeney et al., 2008). Since both these stimulus types involve semantic processing, we avoided them given that semantic or phonological mediation might have been, *a priori*, potential explanations for the correspondence between spoken pseudowords and visual shapes. Additionally, asynchronous audiovisual stimuli are more likely to be perceived as synchronous if the visual element precedes the auditory element (V-A) compared to the reverse (A-V) (Bhat et al., 2015). Since there were equal numbers of V-A and A-V trials in the asynchronous condition, its effectiveness may have been reduced, if some of the V-A trials were actually perceived as synchronous. In order to compensate for this and to test the multisensory integration hypothesis independently of the localizer result, we conducted a further ROI analysis of STS foci derived from a previous study (Werner & Noppeney, 2010) but this showed no significant effects in either of the attended modalities. Nonetheless, pseudoword-shape incongruency-related activations in the right aIPS and SMG overlapped with areas active during the asynchronous condition of the multisensory localizer. The right aIPS incongruency site is close to several IPS foci activated during processing of audiovisual spatial congruency (Sestieri et al., 2006). As noted above, the IPS portions of our parietal incongruency regions overlap with regions implicated in multisensory attention during difficult tasks (Regenbogen et al., 2018). This may be relevant to the contention that the brain is trying harder to integrate these inputs when they are incongruent (Noppeney, 2012). The left MFG incongruency region likely overlaps with a region in inferior frontal cortex responding to audiovisual incongruency for familiar environmental stimuli (Hein et al., 2007) and incongruency between visual primes and auditory targets, whether environmental sounds or spoken words (Noppeney et al., 2008). Since the relevant effects were evoked by multiple types of stimulus incongruency, either of content or of timing, the most plausible explanation for these overlaps is the attention-driven effects cited above, perhaps due to unsuccessful attempts to integrate the audiovisual stimuli. It should be noted that multisensory integration can depend not only on temporal co-occurrence, as in the current localizer task, but also on other types of congruency, such as spatial co-location (Spence, 2013). However, whether spatial co-location is important for integration may be modality- and task-dependent, being less important for some audiovisual (e.g., Regan & Spekreisje, 1977) than visuotactile tasks (e.g., Sambo & Forster, 2009), perhaps especially if, as here, the task is simply to identify the stimulus (Spence, 2013). As noted by Spence (2013, footnote b), auditory stimuli presented via headphones and visual stimuli presented on a screen are different spatial locations and it may be difficult to determine the role of spatial matching in such a set-up. However, participants were both faster and more accurate for congruent than incongruent trials in the ‘attend auditory’ condition, indicating that

the spatially different sources of auditory pseudowords and visual shapes had little effect (in the 'attend visual' condition, participants were faster, but not significantly differentially accurate, for congruent trials).

In the present study, the right SMG incongruency region overlapped with activation during magnitude estimation and was contiguous with the right aIPS magnitude region. The IPS is a classic locus for magnitude processing (Sathian et al., 1999; Eger et al., 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Sokolowski et al., 2017) while the SMG is among regions showing adaptation to magnitude (Piazza et al., 2007) or involved in detecting changes in numerosity (Piazza et al., 2004). Additionally, part of the right POF-precuneus region, whose activation magnitude correlated with RTs during attention to auditory stimuli in the scanner, was close to the right SPG activation evoked by magnitude estimation in the localizer – this region has been implicated in counting visual objects (Sathian et al., 1999). The two pseudowords used in this experiment were drawn, like the shapes, from a much larger set which were empirically rated as sounding more or less rounded or pointed (McCormick et al., 2015), the important point being that participants had no problem in applying magnitude estimation (of roundedness or pointedness) to pseudowords. As noted earlier, sensory features often relate to polar dimensions of magnitude where one end is 'more than' the other (Smith & Sera, 1992), for example brighter vs darker or louder vs quieter, and a localizer targeting such sensory dimensions may have been more relevant to the dimensions of pointedness and roundedness. Nonetheless, there is extensive overlap between brain regions involved in processing number- and sensory-based magnitude (Pinel et al., 2004). However, while we cannot rule out magnitude estimation as a possible mechanism for the observed sound-symbolic incongruency effects, the regions of overlap between the magnitude estimation localizer and the incongruency effects are also implicated in multisensory attentional processes, which may ultimately turn out to be a more important mechanism.

Incongruency effects and attention

We found that, when participants were attending to the auditory pseudowords, several neocortical regions showed greater activity for incongruent, compared to congruent, pairings with visual shapes. Similar effects were not found when participants attended to the visual shapes, which is consistent with the findings that a behavioral congruency effect was absent in the 'attend visual' runs when accuracy was the dependent measure, and was significantly smaller in the 'attend visual' than the 'attend auditory' runs when RT was the dependent measure. These differences as a function of the attended modality may stem from greater unfamiliarity of auditory pseudowords relative to two-dimensional visual shapes. Alternatively, they might reflect timing differences in processing pseudowords compared to visual shapes, since the pseudowords unfold over time, whereas the visual shapes are in evidence from the start of each trial. Consequently, there may be more of an incongruency effect in the 'attend auditory' condition because the concurrent visual stimulus could trigger particular representations

immediately, whereas in the 'attend visual' condition, the concurrent pseudoword may still be playing as the participant prepares a response to the visual stimulus.

The regions demonstrating activations attributable to incongruency were in various parts of parietal cortex bilaterally, and in the left MFG. The left MFG activation likely overlapped with an extensive zone of frontal cortical sensitivity to pseudoword-shape incongruency in a prior study (Peiffer-Smadja, 2010, unpublished thesis, Université Paris Descartes), although this was associated with greater deactivation for congruent than incongruent stimuli in that earlier study The finding of greater cortical activity for incongruent compared to congruent pseudoword-shape pairs may seem puzzling at first glance. One view of congruency manipulations as a test of multisensory integration is that multiple sensory inputs are available to brain regions involved in integration, and that such regions respond more to congruent than incongruent multisensory inputs (Noppeney, 2012). However, Noppeney (2012) has argued that the brain might attempt, unsuccessfully, to integrate incongruent multisensory inputs. If so, regions involved in multisensory integration might, as here, show greater activity for incongruent than congruent stimuli. Moreover, this incongruency effect might be enhanced when, as in the present study, participants selectively attend to one modality while ignoring the other (Noppeney, 2012). For example, when participants attended auditory targets after passively viewing visual primes, greater activity was elicited in several cortical areas (including one that likely overlaps with our left MFG incongruency region) when primes and targets were incongruent compared to congruent (Noppeney et al., 2008). Our left MFG incongruency region also probably overlapped with a focus displaying greater activity for audiovisual stimuli that were familiar but semantically incongruent, compared to unfamiliar arbitrary pairings of abstract images and sounds (Hein et al., 2007). This region also corresponds to part of the domain-general 'multiple demand' attentional system described by Duncan (2013). In the 'attend auditory' incongruent condition, then, what may be happening is that the brain attempts to integrate a stimulus pair but instead detects a mismatch between a rounded pseudoword and a pointed shape or vice versa. On this reasoning, the basis of the observed incongruency-related activations might be greater attentional demand during discrimination of (unfamiliar) auditory pseudowords in the presence of incongruent, compared to congruent, visual shapes. Consistent with this, the incongruency region in the right SMG is similar to a focus exhibiting a preference for novel over familiar stimuli across auditory, visual, and tactile modalities (Downar et al., 2002), and to one in which activation was stronger for unfamiliar abstract audiovisual stimuli compared to familiar, but semantically incongruent, audiovisual stimuli (Hein et al., 2007). Furthermore, the incongruency regions found here in bilateral aIPS and left SPG correspond to regions active during top-down attentional processing when visual targets were cued by semantically incongruent audiovisual stimuli (Mastroberardino et al., 2015), and to bilateral IPS foci (both of which overlap with our incongruency regions) shown to mediate audiovisual interaction during difficult processing of degraded stimuli (Regenbogen et al., 2018). This attentional explanation is especially likely for the subset of incongruency regions, comprising foci in the left IPS/SMG, that also exhibited

correlations between the magnitudes of these neural incongruency effects and the corresponding RT congruency effects during scanning. We acknowledge that these considerations are subject to the limitations of reverse inference, although the limitations tend to be mitigated by our use of task-related reverse inference (Hutzler et al., 2014; McCormick et al., 2018). We suggest that the role of multisensory attention in relation to sound-symbolic crossmodal correspondences merits further research.

Relationship to visual imagery

Sidhu & Pexman (2017) suggest that there may be phonetic elements that reliably co-occur with environmental stimuli. Relatedly, James et al. (2011) showed that the physical shape of objects, either rods or balls, i.e. approximating to pointed or rounded objects, could be reliably inferred from the impact sounds made by dropping these objects onto a hard surface, the shape-selective lateral occipital complex being activated during such inferences. These impact sounds could be mimicked by voiced consonants, either sonorants or plosives, as in 'maluma' or 'bouba' respectively, for ball-like 'rounded' objects; and unvoiced plosive consonants, as in 'takete' or 'kiki', for rod-like 'pointed' objects (Fort et al., 2015). Consistent with this, voiced consonants have been rated as more rounded and unvoiced consonants as more pointed (McCormick et al., 2015). Voiced and unvoiced consonants could also mimic the sound and motion made by round objects rolling, and sharp objects rattling, along a surface. Indeed, the pseudowords 'maluma' and 'bouba' have been associated with the smooth, flowing movements, and 'takete' and 'kiki' with the irregular, jerky movements (Koppensteiner et al., 2016) that we would expect from rounded and pointed objects respectively.

These considerations may be linked to the finding that across participants, increasing preference for object as opposed to spatial imagery was strongly positively correlated with activation magnitudes for the incongruent > congruent contrast in the 'attend auditory' condition. Many of these correlated regions were in visual cortex (in the left IOS, posterior calcarine sulcus/lingual gyrus, and the right MOG), and one was in the left precuneus. This may reflect a greater tendency for object imagers to visualize shapes when associating these with pseudowords. The precuneus (Cavanna & Trimble, 2006), calcarine sulcus, lingual gyrus and MOG (Ganis et al., 2004) are all involved in visual imagery. Another possibility is that, since object imagers tend to integrate structural properties like shape with surface properties like texture (Lacey et al., 2011), the phonetic properties of the pseudowords, for example whether or not consonants are voiced or vowels are rounded, are more strongly bound to the visualized pointed and rounded shapes in these individuals. Interestingly, in contrast to the relatively widespread correlations with imagery preferences, there was little evidence that preference for verbal processing was connected to sound-symbolic associations since correlation of activation magnitudes with OSIVQ verbal scores was limited to a single voxel in right SFS. Further work is required to investigate the potential for individual differences since these might underlie the ability to either guess the

meaning of sound-symbolic words in unknown foreign languages (Kunihira, 1971), or to learn such associations (Nygaard et al., 2009; Revill et al., 2018).

Conclusions

This study is the first to systematically examine the neural mechanisms underlying the sound-symbolic crossmodal correspondence between auditory pseudowords and visual shapes. We used independent localizers and ROI analyses to test a number of *a priori* hypotheses for the neural basis of these correspondences. Overall, evidence for semantic mediation was lacking while that for multisensory integration and magnitude estimation was weak, at best. More robust evidence was found in support of phonological processing as an underlying explanation, and, albeit via task-based reverse inference, multisensory attention emerged as an important potential mechanism, as we also proposed for the crossmodal pitch-elevation correspondence (McCormick et al., 2018). Our findings provide a basis for further research which should seek converging evidence using other methods (for example, multivariate analyses, repetition suppression, or TMS) and also address the relative weights of these different processes. Further, we observed several regions in visual cortex in which activation magnitudes scaled with individual preference for object imagery, thus suggesting a basis for individual differences in processing sound-symbolic associations to be followed up in future research. An important goal for research into crossmodal correspondences in general, as well as sound symbolism in particular, will be to distinguish the roles of high-level cognitive processes, such as phonology and attention, from those of lower-level sensory and perceptual processes.

ACKNOWLEDGMENTS

This work was supported by grants to KS and LN from the National Eye Institute at the NIH (R01EY025978) and the Emory University Research Council; to KMcC from the Emory University Facility for Education and Research in Neuroscience and the Laney Graduate School. Support to KS from the Veterans Administration is also acknowledged. We thank Lawrence Barsalou, Justin Bonny, Daniel Dilks, Evelina Fedorenko, Tami Feng, Harold Gouzoules, Stella Lourenco and Kate Pirog Revill for their advice and assistance.

REFERENCES

Ademollo, F. (2011). *The* Cratylus *of Plato: A Commentary*. Cambridge University Press: Cambridge, UK.

Akita, K. & Tsujimura, N. (2016). Mimetics. In T. Kageyama & H. Kishimoto (Eds.) *Handbook of Japanese Lexicon & Word Formation*, pp133-160. Walter de Gruyter Inc., Boston, USA.

Audacity Team (2012) Audacity v2.0.1 [Computer program]. Retrieved from http://audacity.sourceforge.net/ Audacity ® software is copyright © 1999-2014 Audacity Team.

Beauchamp, M.S. (2005a). Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics*, 5:93-114.

Beauchamp, M.S. (2005b). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, 15:145-153.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E. & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Science USA*, 108:4429-4434.

Ben-Artzi, E. & Marks, L.E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, 57:1151-1162.

Bernstein, I.H. and Edelstein, B.A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, 87:241-247.

Bhat, J., Miller, L.M., Pitt, M.A. & Shahin, A.J. (2015). Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony. *Journal of Neurophysiology*, 113:1437-1450.

Binder, J.R., Westbury, C.F., McKiernan, K.A., Possing, E.T. & Medler, D.A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17:905-917.

Blasi, D.E., Wichmann, S., Hammarström, H., Stadler, P.F. & Christiansen, M.H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences USA*, 113:10818-10823.

Blazhenkova, O. & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied Cognitive Psychology*, 23: 638–663.

Catricalà, M. & Guidi, A. (2015). Onomatopoeias: a new perspective around space, image schemas, and phoneme clusters. *Cognitive Processing*, 16(Suppl 1):S175-S178.

Cavanna, A.E. & Trimble, M.R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129:564-583.

Chen, Y.-C., Huang, P.-C., Woods, A. & Spence, C. (2016). When “bouba” equals “kiki”: cultural commonalities and cultural differences in sound-shape correspondences. *Scientific Reports*, 6:26681, doi:10.1038/srep26681

Crottaz-Herbette, S., & Menon, V. (2006). Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence. *Journal of Cognitive Neuroscience,* 18:766–80.

Dehaene, S., Piazza, M., Pinel, P. & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, 20:487-506.

de Saussure, F. (1916/2009). *Course in General Linguistics*. Open Court Classics: Peru, IL.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language & Linguistics Compass,* 6:654-672.

Downar, J., Crawley, A.P., Mikulis, D.J. & Davis, K.D. (2002). A cortical network sensitive to stimulus salience in a neural behavioral context across multiple sensory modalities. *Journal of Neurophysiology*, 87:615-620.

Duncan, J. (2013). The structure of cognition: attentional episodes in mind and brain. *Neuron,* 80:35-50.

Duvernoy, H.M. (1999). *The Human Brain. Surface, Blood Supply and Three-dimensional Sectional Anatomy*. New York: Springer.

Eger, E., Sterzer, P. Russ, M.O., Giraud, A.-L. & Kleinschmidt, A. (2003). A supramodal number representation in human intraparietal cortex. *Neuron*, 37:719-725.

Eklund, A., Nichols, T.E. & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences USA*, 113:7900-7905.

Erickson, L.C., Heeg, E., Rauschecker, J.P. & Turkeltaub, P.E. (2014). An ALE meta-analysis on the audiovisual integration of speech signals. *Human Brain Mapping*, 35:5587-5605.

Evans, K.K. and Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, 10:6 doi: 10.1167/10.1.6

Fedorenko, E., Behr, M. K., & Kanwisher, N. G. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences USA*, 108:16428-16433.

Fedorenko, E., Duncan, J. & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences USA*, 110:16616-16621.

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

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A. et al. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, 33:636-647.

Fort, M., Martin, A. & Peperkamp, S. (2015). Consonants are more important than vowels in the Bouba-Kiki effect. *Language & Speech*, 58:247-266.

Gallace, A. and Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, 68:1191-1203.

Ganis, G., Thompson, W.L. & Kosslyn, S.M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research*, 20:226-241.

Greenwald, A.G., McGhee, D.E. & Schwarz, J.L.K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality & Social Psychology*, 74:1464-1480.

Hartwigsen, G., Baumgaertner, A., Price, C.J., Koehnke, M., Ulmer, S. & Siebner, H.R. (2010). Phonological decisions require both left and right supramarginal gyri. *Proceedings of the National Academy of Sciences USA*, 107:16494-16499.

Hein, G., Doehrmann, O., Müller, N.G., Kaiser, J., Muckli, L. Naumer, M.J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience*, 27:7881-7887.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press: Oxford, UK.

Jamal, Y., Lacey, S., Nygaard, L. & Sathian, K. (2017). Interactions between auditory elevation, auditory pitch, and visual elevation during multisensory perception. *Multisensory Research*, 30:287-306.

James, T.W., Stevenson, R.A., Kim, S., VanDerKlok, R.M. & James, K.H. (2011). Shape from sound: evidence for a shape operator in the lateral occipital complex. *Neuropsychologia*, 49:1807-1815.

Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. *Linguistics*, 35:379-415.

Knoeferle, K., Li, J., Maggioni, E. & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, 7:5562, doi:10.1038/s41598-017-05965-y

Köhler, W. (1929). *Gestalt Psychology*. Liveright: New York, NY.

Köhler, W. (1947). *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. Liveright: New York, NY.

Koppensteiner, M., Stephan, P. & Jäschke, J.P. (2016). Shaking takete and flowing maluma: nonsense words are associated motion patterns. *PLoS ONE*, 11:e0150610, doi:10.1371/journal.pone.0150610

Ković, V., Plunkett, K. & Westermann, G. (2010). The shape of words in the brain. *Cognition*, 114:19-28.

Kunihira, S. (1971). Effects of the expressive voice on phonetic symbolism. *Journal of Verbal Learning and Behavior*, 10:427-429.

Lacadie, C.M., Fulbright, R.K., Constable, R.T. & Papademetris, X. (2008). More accurate Talairach coordinates for neuroimaging using nonlinear registration. NeuroImage, 42:717-725.

Lacey, S., Lin, J.B. & Sathian, K. (2011). Object and spatial imagery dimensions in visuo-haptic representations. *Experimental Brain Research*, 213(2-3), 267-273.

Lacey, S., Martinez, M.O., McCormick, K. and Sathian, K. (2016). Synesthesia strengthens sound-symbolic cross-modal correspondences. *European Journal of Neuroscience*, 44:2716-2721.

Lockwood, G., Hagoort, P. & Dingemanse, M. (2016). How iconicity helps people learn new words: neural correlates and individual differences in sound-symbolic bootstrapping. *Collabra*, 2:7, doi: 10.1525/Collabra.42

Lourenco, S.F., Bonny, J.W., Fernandez, E. P. & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences USA*, 109:18737-18742.

Lourenco, S. F., & Longo, M. R. (2011). Origins and development of generalized magnitude representation. In S. Dehaene and E. Brannon (Eds.), *Space, Time, and Number in the Brain: Searching for the Foundations of Mathematical Thought.* (pp. 225-244). Elsevier.

McCormick, K. (2015). [Sound to visual shape mappings]. Unpublished raw data.

McCormick, K., Kim, J.Y., List, S. & Nygaard, L.C. (2015). Sound to meaning mappings in the bouba-kiki effect. *Proceedings 37th Annual Meeting Cognitive Science Society*, 1565-1570.

McCormick, K., Lacey, S., Stilla, R., Nygaard, L.C. & Sathian, K. (2018). Neural basis of the crossmodal correspondence between auditory pitch and visuospatial elevation. *Neuropsychologia*, 112:19-30.

Marchant, J.L., Ruff, C.C. & Driver, J. (2012). Audiovisual asynchrony enhances BOLD responses in a brain network including multisensory STS while also enhancing target-detection performance for both modalities. *Human Brain Mapping*, 33:1212-1224.

Mastroberardino, S., Santangelo, V. & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, 9:45, doi:10.3389/fnint.2015.00045

Mikl, M., Mareček, R., Hluštik, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M. & Krupa, P. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, 26:490-503.

Monaghan, P., Mattock, K. & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 38:1152-1164.

Nielsen, A. & Rendall, D. (2011). The sound of round: evaluating the sound-symbolic role of consonants in the classic *Takete-Maluma* phenomenon. *Canadian Journal of Experimental Psychology*, 65:115-124.

Noesselt, T., Bergmann, D., Heinze, H.-J., Münte, T. & Spence, C. (2012). Coding of multisensory temporal patterns in human superior temporal sulcus. *Frontiers in Integrative Neuroscience*, 6:64, doi: 10.3389/fnint.2012.00064

Noppeney, U. (2012). Characterization of multisensory integration with fMRI. In M.M. Murray & M.T. Wallace (Eds.), *The Neural Bases of Multisensory Processes*. (pp233-252). CRC Press.

Noppeney, U., Josephs, O., Hocking, J., Price, C.J. & Friston, K.J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex*, 18:598-609.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, 112:181–186.

Oberhuber, M., Hope, T.M.H., Seghier, M.L., Jones, O.P., Prejawa, S., Green, D.W. & Price, C.J. (2016). Four functionally distinct regions in the left supramarginal gyrus support word processing. *Cerebral Cortex*, 26:4212-4226.

Ozturk, O., Krehm, M. & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114:173-186.

Parise, C.V., Knorre, K. and Ernst, M.O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences USA*, 111:6104-6108.

Parise, C.V. & Spence, C. (2012). Audiovisual cross-modal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, 220:319-333.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D. & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44:547-555.

Piazza, M., Pinel, P., Le Bihan, D. & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, 53:293-305.

Pinel, P., Piazza, M., Le Bihan, D. & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, 41:983-993.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. Harper Collins.

Price, C.J., Moore, C.J., Humphreys, G.W. & Wise, R.J.S. (1997). Segregating semantic from phonological processes during reading. *Journal of Cognitive Neuroscience*, 9:727-733.

Raczkowski, D., Kalat, J.W. & Nebes, R. (1974). Reliability and validity of some handedness questionnaire items. *Neuropsychologia*, 12:43-47.

Regan, D. & Spekreisje, H. (1977). Auditory-visual interactions and the correspondence between perceived auditory space and perceived visual space. *Perception*, 6:133-138.

Regenbogen, C., Seubert, J., Johannson, E. Finkelmeyer, A., Andersson, P. & Lundström, J.N. (2018). The intraparietal sulcus governs multisensory integration of audiovisual information based on task difficulty. *Human Brain Mapping*, 39:1313-1326.

Revill, K.P., Namy, L.L., DeFife, L.C. & Nygaard, L.C. (2014). Cross-linguistic sound symbolism and crossmodal correspondence: evidence from fMRI and DTI. *Brain & Language*, 128:18-24.

Revill, K. P., Namy, L. L., & Nygaard, L. C. (2018). Eye movements reveal persistent sensitivity to sound symbolism during word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44:680-698.

Sambo, C.F. & Forster, B. (2009). An ERP investigation on visuotactile interactions inn peripersonal and extrapersonal space: evidence for the spatial rule. *Journal of Cognitive Neuroscience*, 21:1550-1559.

Sathian, K., Simon, T.J., Peterson, S., Patel, G.A., Hoffman, J.M. & Grafton, S.T. (1999). Neural evidence linking visual object enumeration and attention. *Journal of Cognitive Neuroscience*, 11:36-51.

Schmidtke, D.S., Conrad, M. & Jacobs, A.M. (2014). Phonological iconicity. *Frontiers in Psychology*, 5:80, doi: 10.3389/fpsyg.2014.00080

Sestieri, C., Di Matteo, R., Ferretti, A., Del Gratta, C., Caulo, M., Tartaro, A., Belardinelli, M.O. & Romani, G.L. (2006). "What" versus "where" in the audiovisual domain: an fMRI study. *NeuroImage*, 33:672-680.

Sidhu, D.M. & Pexman, P.M. (2017). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, in press, doi:10.3758/s13423-017-1361-1

Smith, L.B. & Sera M.D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, 24:99-142.

Sokolowksi, H.M., Fias, W., Ononye, C.B. & Ansari, D. (2017). Are numbers grounded in a general magnitude processing system? A functional neuroimaging meta-analysis. *Neuropsychologia*, 105:50-69.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception & Psychophysics*, 73:971-95.

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences*, 1296:31-49.

Stevenson, R.A., Altieri, N.A., Kim, S., Pisoni, D.B. & James, T.W. (2010). Neural processing of asynchronous audiovisual speech perception. *NeuroImage*, 49:3308-3318.

Styles, S.J. & Gawne, L. (2017). When does maluma/takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception*, doi:10.1177/2041669517724807

Sučević, J., Savić, A.M., Popović, M.B., Styles, S.J. & Ković, V. (2015). Balloons and bavoons versus spikes and shikes: ERPs reveal shared neural processes for shape-sound-meaning congruence in words, and shape-sound congruence in pseudowords. *Brain & Language*, 145/146:11-22.

Talairach, J. & Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme Medical Publishers; New York.

Thompson, P.D. & Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *Quarterly Journal of Experimental Psychology*, 64:2392-2404.

van Atteveldt, N.M., Formisano, E., Blomert, L. & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex*, 17:962-974.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, 7:483-488.

Werner, S. & Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex*, 20:1829-1842.

White, T., O'Leary, D., Magnotta, V., Arndt, S., Flaum, M. & Andreasen, N.C. (2001). Anatomic and functional variability: The effects of filter size in group fMRI data analysis. *NeuroImage*, 13:577-588.

Wilson, L.B., Tregellas, J.R., Slason, E., Pasko, B.E. & Rojas, D.C. (2011). Implicit phonological priming during visual word recognition. *NeuroImage*, 55:724-731.

Woo, C.-W., Krishnan, A. & Wager, T.D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, 91:412-419.

Abbreviations

Directional: a, anterior; front, frontal; i, inferior; lat, lateral; m, mid; med, medial; p, posterior; s, superior; v, ventral.

Anatomical: AG, angular gyrus; AOS, anterior occipital sulcus; calcS, calcarine sulcus; CeS, central sulcus; cingG, cingulate gyrus; cingS, cingulate sulcus; collatS, collateral sulcus; FG, fusiform gyrus; fo, frontal operculum; InfOS, inferior occipital sulcus; Ins, insula; IOG, inferior occipital gyrus; IOS, intra-occipital sulcus; IPS, intraparietal sulcus; ITG, inferior temporal gyrus; ITS, inferior temporal sulcus; LG, lingual gyrus; MFG, middle frontal gyrus; MOG, middle occipital gyrus; OrbG, orbital gyrus; poCG, post central gyrus; po, pars opercularis; poCS, post central sulcus; POF, parieto-occipital fissure; preCS, precentral sulcus; preCG, precentral gyrus; precun, precuneus; preSMA, presupplementary motor area; pt, pars triangularis; SFG, superior frontal gyrus; SFS, superior frontal sulcus; SMG, supramarginal gyrus; SOG, superior occipital gyrus; SPG, superior parietal gyrus; STS, superior temporal sulcus. All other abbreviations are as in the main text.

Table 1 Localizer activations: semantic (a,b), multisensory integration (c,d), and magnitude (e); all within cortical mask, voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster thresholds = semantic, 9 voxels; multisensory integration, 8 voxels; magnitude, 7 voxels; x,y,z Talairach coordinates for centers of gravity.

	Region	**x**	**y**	**z**
(a) Sentences > pseudowords	R a STG - a STS - mid STS - p STS	49	-6	-8
	L preCG – preCS - MFG	-44	-5	48
	L MOG	-40	-69	21
	L a STG - a STS - mid STS - p STS - p STG	-51	-19	-3
(b) Pseudowords > sentences	R SFG - SFS - MFG	29	19	47
	R SFS - SFG - MFG - IFS - IFG	32	49	15
	R med SFG	8	33	31
	R p cingS - p cingG	4	-31	36
	R precun - POF	11	-62	29
	R AG – mid IPS	46	-57	35
	L med SFG	-2	31	32
	L p cingS - p cingG	-4	-29	35
	L MFG - SFS	-31	28	38
	L IFS - IFG - lat OrbG - a OrbG	-31	48	8
	L SMG - AG	-44	-52	41
	L SPG - IPS - POF	-12	-68	31
	L MOG - InfOS	-36	-79	-3
(c) Synchronous > asynchronous	R a calcS - POF - cuneus	2	-60	15
	L a calcS - POF - cuneus - LG -p cingG	-4	-58	15
(d) Asynchronous > synchronous	R SMG – a IPS – mid IPS - AG	38	-47	42
	R IFG	53	11	16
(e) Magnitude > control	R SMG	42	-40	47
	R SPG - av IPS	18	-63	42
	R MOG - IOS - SOG	25	-86	8

Table 2 Incongruency effects: pseudoword-shape incongruency-related activations in the 'attend auditory' condition within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$ cluster threshold 5 voxels); x,y,z, Talairach coordinates for centers of gravity.

Region	**x**	**y**	**z**
R aIPS[1]	33	-40	42
R SMG – poCS[1,2]	45	-38	45
L SPG	-16	-61	58
L mid IPS - SMG – aIPS[3]	-36	-47	44
L MFG[3]	-42	30	28

[1] Overlaps with the asynchronous R aIPS and SMG in Table 1d.
[2] Overlaps with the R SMG and contiguous with the R aIPS magnitude foci in Table 1e.
[3] Overlaps with the semantic control L SMG and MFG in Table 1b.

Table 3 Correlations within cortical mask (r-map threshold p < .001, cluster-corrected p < .05 (except (b), FDR corrected q < .05), cluster thresholds = (a, c) 5 voxels, (d) 4 voxels; x,y,z, Talairach coordinates for centers of gravity.

Region	x	y	z	Mean r
(a) Attend visual (C > I) – RT congruency effect				
L SMG[1]	-50	-45	35	-.72
(b) Attend visual (C > I) – OSIVQ verbal score				
R SFS[2]	27	8	49	.70
(c) Attend auditory (I > C) – RT congruency effect				
R preSMA	3	5	52	.73
R POF – precun[3]	12	-66	37	.73
R IFS - IFG	32	12	28	.72
R a Ins - fo	34	25	7	.74
L preSMA	-6	5	47	.72
L pIPS - midIPS – AG[4]	-27	-57	41	.73
L aIPS – SMG[5]	-35	-45	36	.71
(d) Attend auditory (I > C) – OSdiff score				
R MOG	44	-61	3	.73
L SMA	-5	-5	63	.71
L precun	-6	-57	52	.74
L IOS	-22	-79	20	.74
L p calcS - LG	-10	-84	-6	.72

[1] Overlaps with the semantic control L SMG in Table 1b.

[2] Contiguous with the semantic control R SFS/MFG in Table 1b.

[3] Contiguous with the semantic control R precun – POF in Table 1b.

[4] Contiguous with the semantic control L AG/IPS in Table 1b.

[5] Overlaps with the incongruency effect L a IPS/SMG in Table 2 and contiguous with the semantic control L SMG in Table 1b.

FIGURE LEGENDS

Figure 1 Example stimuli for (a) the sound-symbolic word-shape correspondence task and the localizer tasks: (b) semantic, (c) multisensory integration, (d) magnitude.

Figure 2 Semantic localizer within cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 9 voxels). Contrast of sentences > pseudowords (orange) reveals semantic network (Table 1a); pseudowords > sentences (olive) likely reflect phonological processing (Table 1b).

Figure 3 Multisensory integration localizer within cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 8 voxels). Contrast of synchronous > asynchronous (turquoise) reveals putative integration network (Table 1c); asynchronous > synchronous (yellow: Table 1d).

Figure 4 Magnitude localizer within cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 7 voxels). Contrast of magnitude estimation > control reveals magnitude network (Table 1e).

Figure 5 Sound-symbolic incongruency effects: pseudoword-shape incongruency-related activations in the attend auditory condition (Table 2).

Figure 6 Representative time-course curves for regions showing sound-symbolic incongruency effects in (A) the incongruent > congruent contrast and (B-D) contrasts from the localizers with which they shared an overlap zone.

Figure 7 Sound-symbolic incongruency effects (blue) in relation to functional localizer activations for semantics (orange), phonology (olive), multisensory integration (turquoise), and magnitude processing (green). Circles indicate areas of overlap or contiguity between incongruency effects and localizer conditions (see notes to Table 2).

Figure 8 Cortical regions showing correlations between activation magnitudes for the C > I contrast in the 'attend visual' condition and the magnitude of the RT congruency effect (brown: Table 3a) and scores on the verbal sub-scale of the OSIVQ (blue-gray: Table 3b); and between activation magnitudes for the I > C contrast in the 'attend auditory condition and the magnitude of the RT congruency effect (red: Table 3c) and OSdiff scores (magenta: Table 3d) – higher OSdiff scores indicate stronger preference for object, rather than spatial, imagery. See notes to Table 3 for relationships to pseudoword-shape incongruency regions and functional localizer activations.

Figure 1

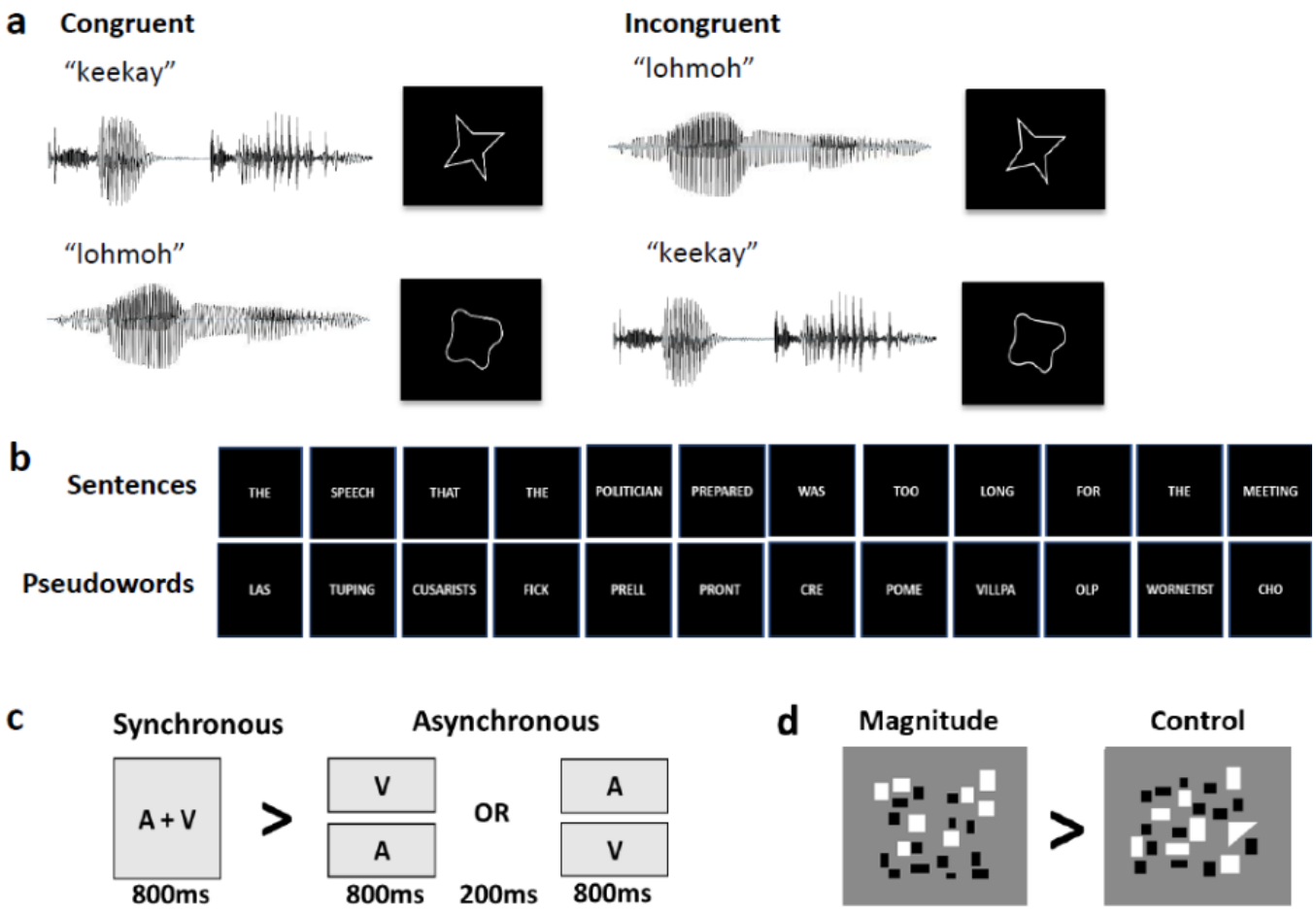

Figure 2

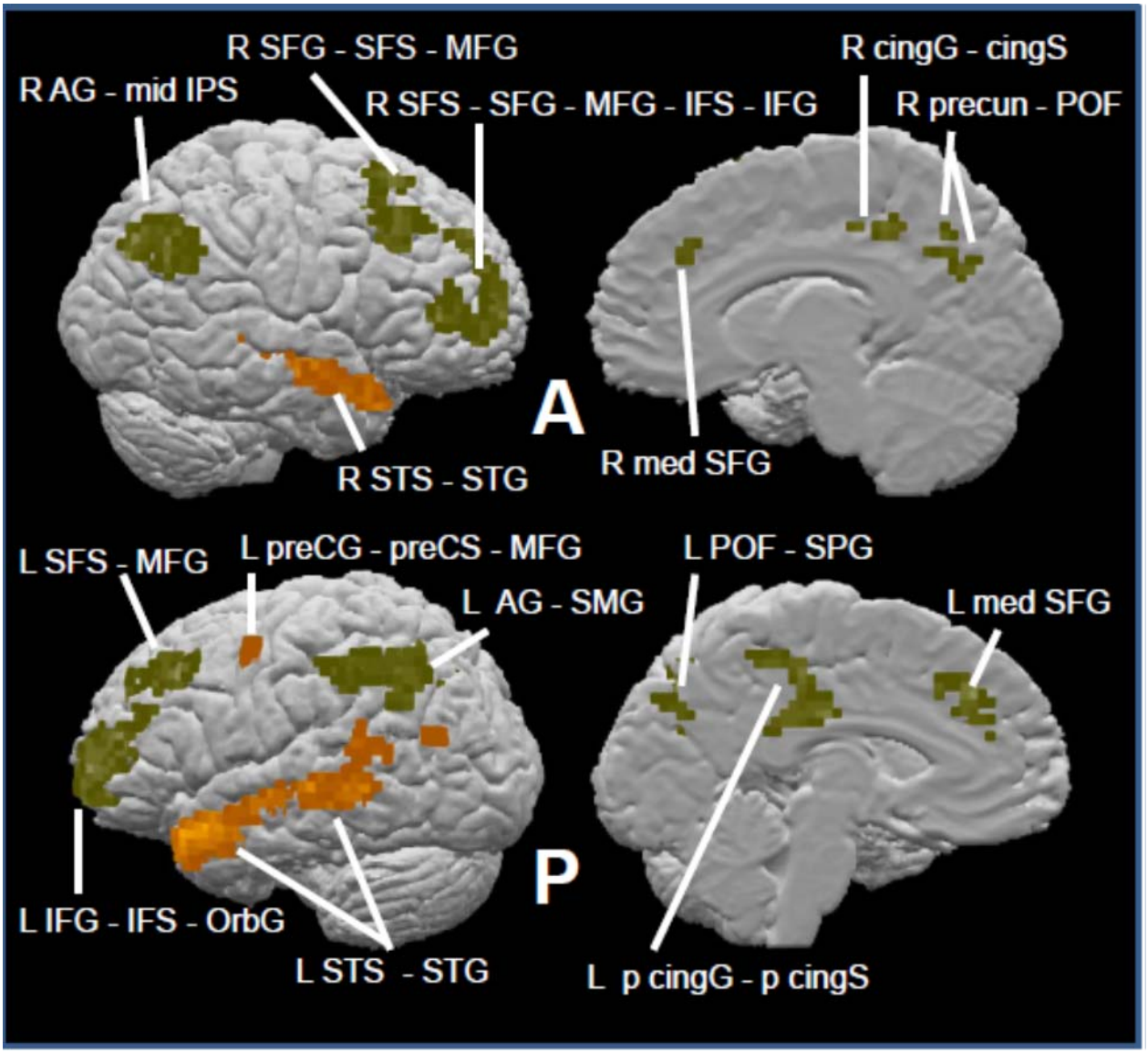

Figure 3

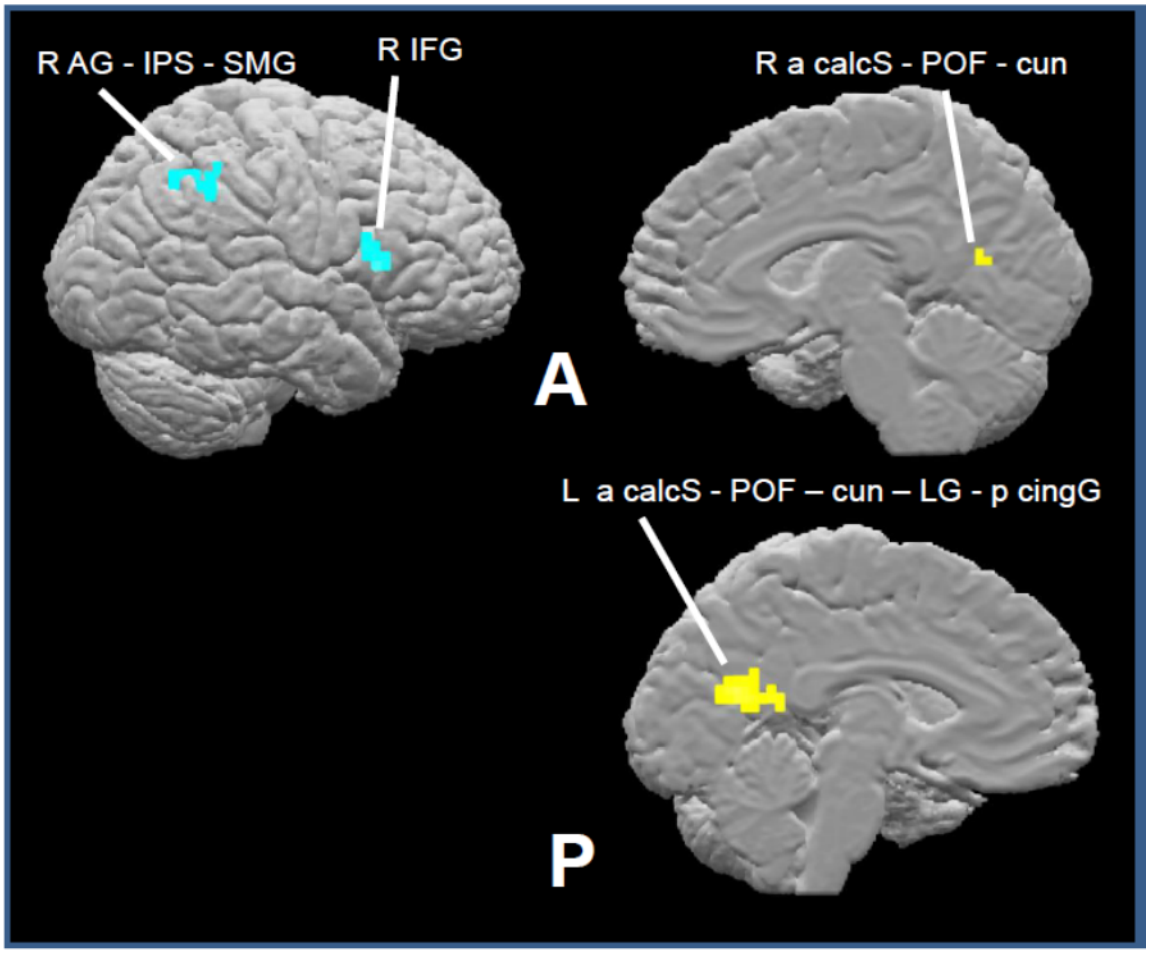

Figure 4

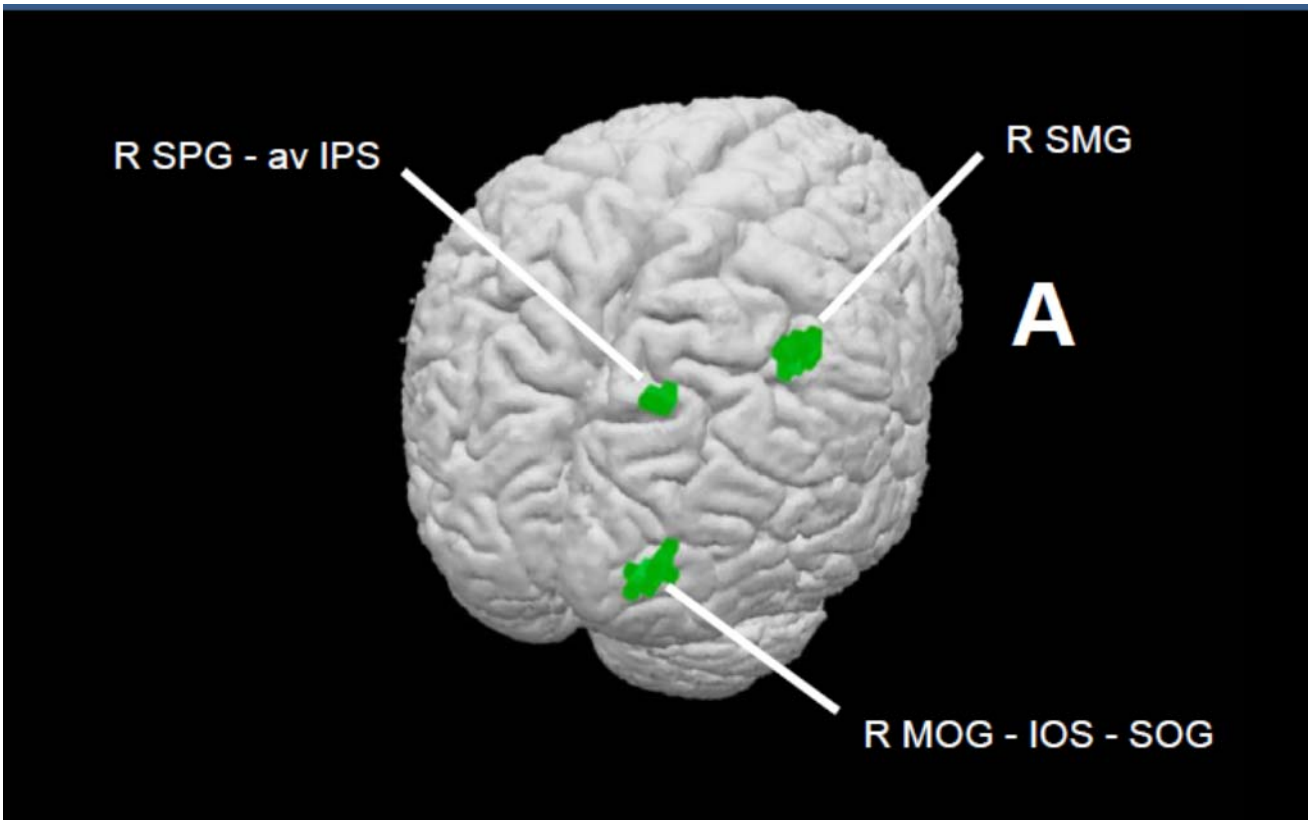

Figure 5

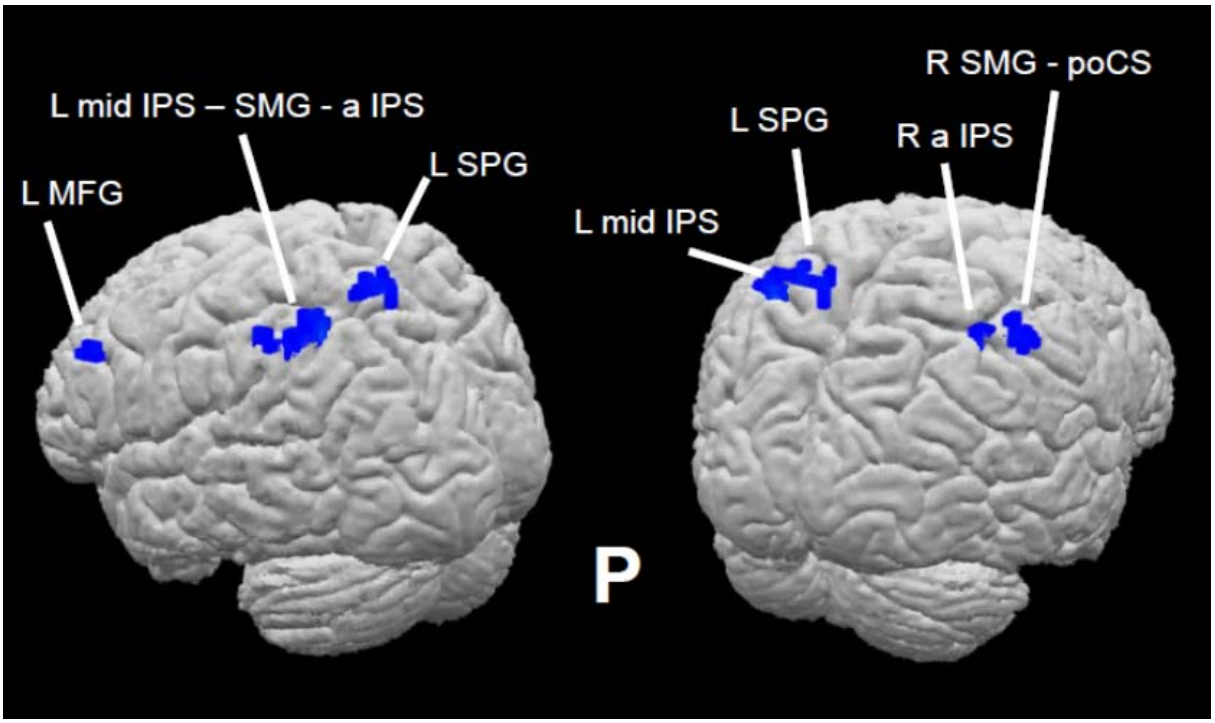

Figure 6

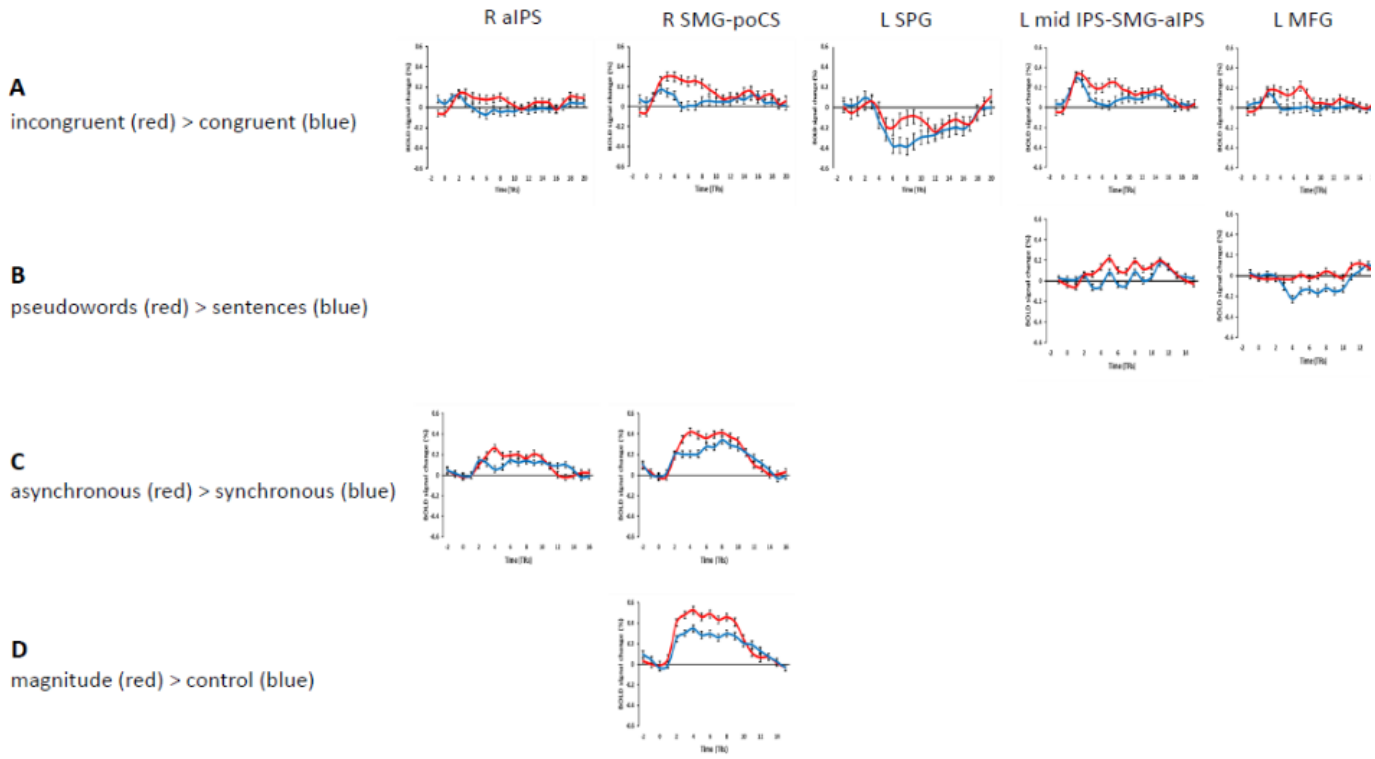

Figure 7

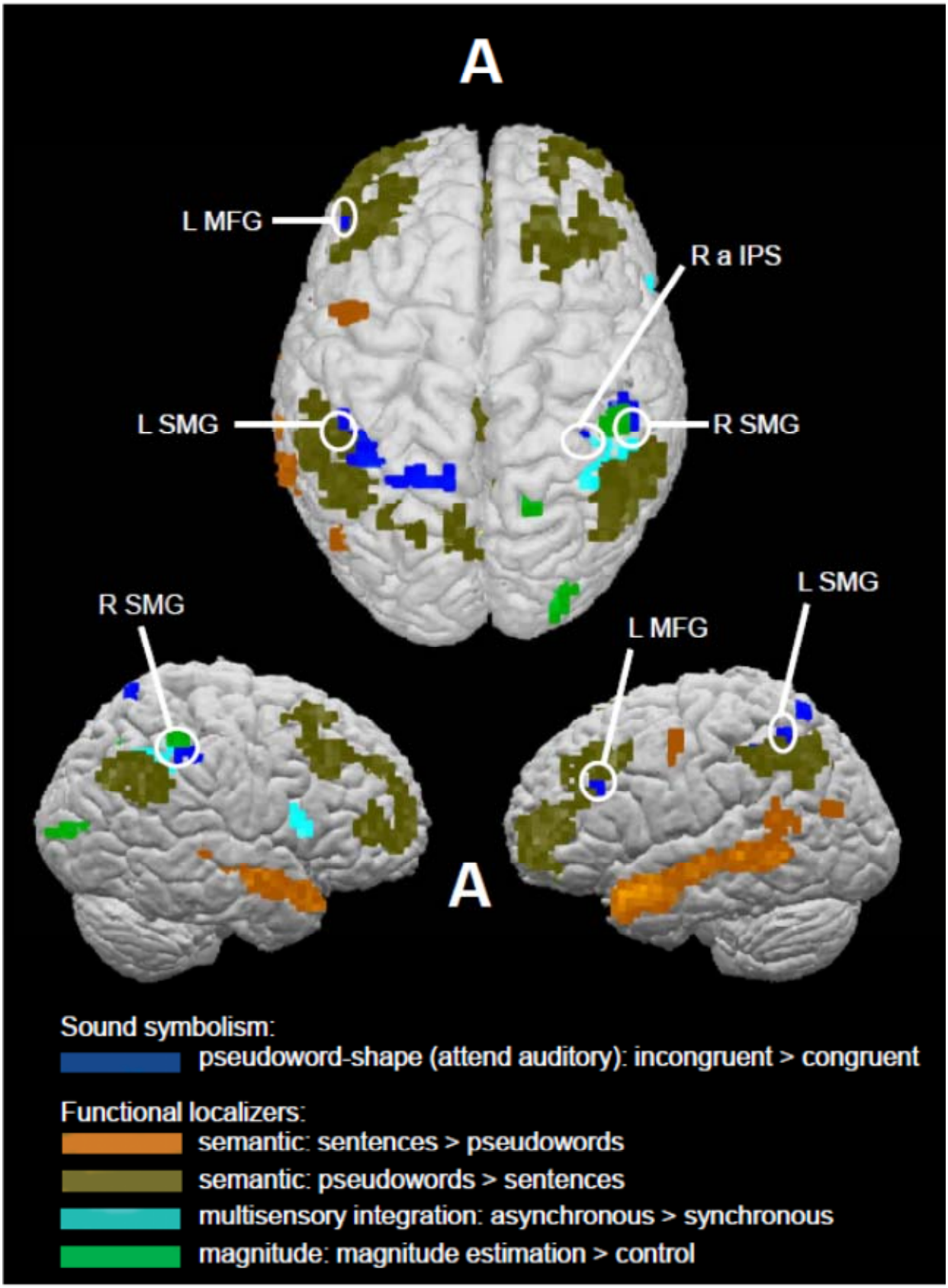

Figure 8