HHS Public Access

Author manuscript

Neuropsychologia. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

Neuropsychologia. 2019 September ; 132: 107132. doi:10.1016/j.neuropsychologia.2019.107132.

Speech-accompanying gestures are not processed by the language-processing mechanisms

Olessia Jouravlev[1,2], **David Zheng**[3], **Zuzanna Balewski**[1], **Alvince Le Arnz Pongos**[1], **Zena Levan**[4], **Susan Goldin-Meadow**[4], **Evelina Fedorenko**[1,5,6]

[1]Massachusetts Institute of Technology, Cambridge, MA 02139

[2]Carleton University, Ottawa, ON K1S 5B6, Canada.

[3]Princeton University, Princeton, NJ 08544

[4]University of Chicago, Chicago, IL, 60637

[5]McGovern Institute for Brain Research, Cambridge, MA, 02139

[6]Massachusetts General Hospital, Boston, MA, 02114

Abstract

Speech-accompanying gestures constitute one information channel during communication. Some have argued that processing gestures engages the brain regions that support language comprehension. However, studies that have been used as evidence for shared mechanisms suffer from one or more of the following limitations: they a) have not directly compared activations for gesture and language processing in the same study and relied on the fallacious reverse inference (Poldrack, 2006) for interpretation, b) relied on traditional group analyses, which are bound to overestimate overlap (e.g., Nieto-Castañon & Fedorenko, 2012), c) failed to directly compare the magnitudes of response (e.g., Chen et al., 2017), and d) focused on gestures that may have activated the corresponding linguistic representations (e.g., “emblems”). To circumvent these limitations, we used fMRI to examine responses to gesture processing in language regions defined functionally in individual participants (e.g., Fedorenko et al., 2010), including directly comparing effect sizes, and covering a broad range of spontaneously generated co-speech gestures. Whenever speech was present, language regions responded robustly (and to a similar degree regardless of whether the video contained gestures or grooming movements). In contrast, and critically, responses in the language regions were low – at or slightly above the fixation baseline – when silent videos were processed (again, regardless of whether they contained gestures or grooming movements). Brain regions outside of the language network, including some in close proximity to its regions, differentiated between gestures and grooming movements, ruling out the possibility that the gesture/grooming manipulation was too subtle. Behavioral studies on the critical video

Address for Correspondence: Olessia Jouravlev (olessiaj@mit.edu) or Ev Fedorenko (evelina9@mit.edu); Massachusetts Institute of Technology, Brain & Cognitive Sciences Department, 43 Vassar Street, Building 46, Room 3037, Cambridge, MA 02139.

Conflict of interest: The authors declare no competing financial interests.

materials further showed robust differentiation between the gesture and grooming conditions. In summary, contra prior claims, language-processing regions do not respond to co-speech gestures in the absence of speech, suggesting that these regions are selectively driven by linguistic input (e.g., Fedorenko et al., 2011). Although co-speech gestures are uncontroversially important in communication, they appear to be processed in brain regions distinct from those that support language comprehension, similar to other extra-linguistic communicative signals, like facial expressions and prosody.

Keywords

Co-speech gestures; Language Network; Functional Specificity; Multiple Demand (MD) Network; Communication

Introduction

Research on language comprehension has been dominated by investigations of how *linguistic* information is processed. However, in naturalistic communicative settings, in addition to the content of the linguistic signal, comprehenders use a wealth of other information carried by prosody (e.g., Cutler, Dahan, & Van Donselaar, 1997), voice quality (e.g., Gobl & Chasaide, 2003), non-verbal vocalizations (e.g., Campbell, 2007), facial expressions (e.g., Harper, Wiens, & Matarazzo, 1978), eye gaze (e.g., Kendon, 1967), body posture (e.g., Müller et al., 2013), and spontaneous hand and arm movements, or gestures (e.g., Abner, Cooperrider, & Goldin-Meadow, 2015; Goldin-Meadow, 2003; McNeill, 1992, 2000; Novack & Goldin-Meadow, 2016). The nature of the mechanisms that support the processing of extra-linguistic information during communication remains debated (Buck, 1984; Buck & VanLear, 2002, Hage & Nieder, 2016; Rauschecker, 2018).

One important question is whether extra-linguistic signals are processed by the same cognitive and neural mechanisms that process linguistic content. A priori, based on theoretical considerations, it is possible to argue for either overlapping or distinct mechanisms. On the one hand, both linguistic and extra-linguistic signals carry socially- and communicatively-relevant information. If (any of) the brain regions that respond robustly to language (e.g., Binder et al., 1997; Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010) respond to the communicative property of linguistic stimuli, we would expect these regions to also respond strongly to non-linguistic stimuli that carry communicatively-relevant information. On the other hand, linguistic and extra-linguistic signals are characterized by distinct properties, and thus plausibly place distinct computational demands on the mind and brain. In particular, language relies on a set of conventionalized form-meaning mappings, both at the level of individual words and constructions (e.g., Goldberg, 1995). Furthermore, language is inherently compositional: by combining individual words and constructions, we create new complex meanings (e.g., von Humboldt, 1836/1999). In contrast, although extra-linguistic signals carry information, they are not typically associated with conventionalized meanings, nor do they compose in the same way as words do. If the language brain regions respond to feature(s) of linguistic stimuli that are *not* shared with non-linguistic communicative signals, we would expect low

responses in these areas to anything but language (e.g., Fedorenko, Behr, & Kanwisher, 2011).

So, do the same mechanisms support the processing of linguistic vs. extra-linguistic signals? For some extra-linguistic signals, the answer appears to be a clear 'no'. For example, the processing of emotional prosody (e.g., Ross, 1981; Weintraub, Mesulam, & Kramer, 1981), facial expressions (e.g., Adolphs, 2002; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011), and eye gaze (e.g., Itier & Batty, 2009) have long been linked to the right hemisphere, in contrast to the left-hemisphere dominance for language comprehension (see Pritchett, Hoeflin, Koldewyn, Dechter, & Fedorenko, 2018, for direct fMRI evidence that language-responsive brain regions do not respond to observing others' eye, mouth, or face movements). However, for other extra-linguistic signals, like gestures, the answer is less clear. A number of prior brain imaging studies of gestures have reported activations within the broad anatomical areas historically implicated in language processing – left posterior temporal and inferior frontal cortices (e.g., Dick, Goldin-Meadow, Hasson, Skipper, & Small, 2009; Holle, Gunter, Rueschemeyer, Hennenlotter, & Iacoboni, 2008; Holler et al., 2015; Hubbard, Wilson, Callan, & Dapretto, 2009; Weisberg, Hubbard, & Emmorey, 2016; for reviews, see Willems & Hagoort, 2007; Marstaller & Burianová, 2014; Yang, Andric, & Mathew, 2015). But other gesture studies have failed to observe effects in these brain areas (e.g., Willems, Ozyürek, & Hagoort, 2009), and at least some studies have reported intact gesture perception (e.g., Eggenberger et al., 2016) and production (de Beer et al., 2017) in aphasia. These apparent inconsistencies in the prior literature may, at least in part, be due to methodological considerations. In particular, most prior neuroimaging studies have suffered from one or more of the following limitations.

First, interpretation of the activation peaks/clusters and comparisons across studies (e.g., comparing a cluster of activation observed for a gesture manipulation to prior studies on language processing) are often made at the level of *coarse macro-anatomical landmarks* (e.g., the STS, MTG, or IFG). Reliance on such broad anatomical definitions can be misleading because functionally distinct regions often lie in close proximity to each other within the same anatomical structures. Indeed, such functional heterogeneity has been reported for both left lateral temporal structures, like the STS (e.g., Deen, Koldewyn, Kanwisher, & Saxe, 2015), and left inferior frontal structures (e.g., Fedorenko, Duncan, & Kanwisher, 2012).

Second, studies that directly contrast language and gesture processing either examine overlap at the *group level* (e.g., Andric et al., 2013; Dick, Goldin-Meadow, Solodkin, & Small, 2012, Dick, Mok, Beharelle, Goldin-Meadow, & Small, 2014; Green et al., 2009; He et al., 2015; cf. Redcay, Velnoskey, & Rowe, 2016, for a recent exception), and/or fail to directly compare the *magnitudes of response.* Overlap between conditions is likely to be overestimated in traditional group analyses (e.g., Nieto-Castañon & Fedorenko, 2012) given the well-documented inter-individual variability in the locations of functional regions especially pronounced in the association cortices (e.g., Frost & Goebel, 2011; Tahmasebi et al., 2012). And an effect size comparison is critical for interpreting the observed overlap (e.g., Sullivan & Feinn, 2012; see Chen, Taylor, & Cox, 2017, for a recent discussion of this issue in the context of fMRI research). As an example, consider the case of a well-

characterized region in the ventral visual stream – the fusiform face area, or the FFA (Kanwisher, McDermott, & Chun, 1997). Every voxel in this region responds highly significantly to both faces and non-face objects, yet the response to faces is 2-3 times stronger than the response to non-face objects (and discovering this selectivity had laid a core foundation for mechanistic-level accounts of face recognition; e.g., Chang & Tsao, 2017). Simply reporting significance (*t/p*) maps for the two conditions, or the overlap map, would therefore be highly misleading, as it would completely obscure the large and meaningful difference in the magnitude of the actual neuronal response we are attempting to measure with fMRI.

And finally, some studies have specifically focused on *symbolic gestures* or "emblems" (e.g., wave, hold out hand for a shake, etc.; Andric et al., 2013; Enrici et al., 2011; Xu et al., 2009; Redcay et al., 2016; Papeo et al., 2019). Given that emblems appear to function like words, processing an emblem may activate the linguistic representation of the word(s) associated with that emblem. Thus, the overlap may be explained by engagement of linguistic resources during gesture processing, rather than by computational demands shared by the two. This criticism also applies to studies that have examined cases where gestures clarify the linguistic message (e.g., producing a 'playing guitar' gesture when saying 'He plays an instrument'; Demir-Lira et al., 2018; Dick et al., 2014; Kircher et al., 2009; Straube, Green, Bromberger, & Kircher, 2011; Willems, Ozyürek, & Hagoort, 2009). In such cases, the gesture plausibly activates the relevant lexical representation (e.g., "guitar" in the example above), so a stronger response in the language areas for these cases, compared to scenarios where the gesture does not add extra information, are difficult to interpret as indexing overlap in the processing mechanisms.

To definitively address the question of whether language-processing brain regions support gesture processing – which is needed for a more comprehensive understanding of the cognitive and neural substrates of linguistic vs. non-linguistic communication – we examined responses of language areas to gesture processing (i) at the level of individual participants (e.g., Fedorenko et al., 2010), including (ii) a direct comparison of effect sizes, and (iii) covering a broad range of spontaneously generated co-speech gestures.

Methods

General approach and key predictions.

We leveraged our knowledge about the core language network – a set of frontal and temporal brain areas that selectively support high-level language processing (Fedorenko et al., 2011), including lexico-semantic and combinatorial processing (Fedorenko et al., 2010, Fedorenko, Nieto-Castaon, & Kanwisher, 2012a; Fedorenko et al., 2018; Bautista & Wilson, 2016) – to probe the responses of these regions to speech-accompanying gestures.

An important consideration is what to use as a control for gestures. Prior fMRI investigations of gesture processing have used diverse control conditions, including low-level baselines like rest or fixation (e.g., Josse, Joseph, Bertasi, & Giraud, 2012; Nagels, Chatterjee, Kircher, & Straube, 2013), speech without gestures (e.g., Green et al., 2009; Kircher et al., 2009), meaningless movements (e.g., Straube, Green, Weis, & Kircher, 2012),

and grooming movements, like scratching or fixing one's hair (e.g., Dick et al., 2009, 2012; Holle et al., 2008; Skipper, Goldin-Meadow, Nusbaum, & Small, 2007). We chose grooming movements as a control condition for gestures because both (a) are rich and dynamic visual stimuli, (b) contain familiar biological movements (see Behavioral Study 2 below), and (c) carry some information. A key difference between the two is that gestures have been shown to relate to the speech signal semantically and temporally (Kendon, 1980; McNeill, 1992), but there is no evidence that grooming movements show the same relation to speech.

The experiment consisted of five conditions: two silent video conditions (with gestures vs. grooming movements), two video conditions where speech was present (again, with gestures vs. grooming movements), and a condition where the speech signal was presented without any visual input, which is expected to elicit a robust response across the language network (e.g., Scott, Gallée, & Fedorenko, 2016).

The *silent video conditions* are critical for the primary research question we ask here. In particular, if the language regions respond to communicatively-relevant signals regardless of whether they are linguistic or not, then silent gesture videos should elicit a strong response. This response should be stronger than the control, grooming videos, condition, and it should be similarly strong to the speech-only condition (Figure 1a). (Note that on this account, there may or may not be a difference between the gesture and grooming conditions in the presence of speech, depending on whether speech alone is sufficient for driving these regions to their maximal response capacity.) If, instead, the language regions are selective for linguistic content (e.g., Fedorenko et al., 2011), then the response to silent gesture videos should be low and no different from the silent grooming videos condition (Figure 1b).

The *video conditions with speech* allow us to test three additional hypotheses: (1) The language regions respond to the total amount of communicatively-relevant information (regardless of potential redundancies in the information in the different signals; cf. Dick et al., 2014); (2) The language regions respond to the demands associated with the need to integrate communicatively-relevant information from different signals (e.g., Demir-Lira et al., 2018; Dick et al., 2009; 2012); and (3) The language regions respond to communicatively-relevant signals, but gestures only become communicatively-relevant in the presence of speech (cf. Behavioral Studies 3 and 4). All three of these hypotheses predict that adding gestures to a linguistic signal should lead to an increase in the neural response relative to the speech-only condition because gestures carry communicatively-relevant information. This should not be the case for adding grooming movements to a linguistic signal given that grooming movements are not, or at least are a lot less, communicatively-relevant (see Behavioral study 3 below) (Figures 1c–d). The first hypothesis further predicts a stronger response to silent gesture videos (given that they contain communicatively-relevant information; see Behavioral studies 3 and 4 below) than silent grooming videos (Figure 1c). The second and third hypotheses predict a low response to the two silent video conditions (Figure 1d).

Participants.

Seventeen individuals (mean age 23, 9 females), all right-handed (as determined by the Edinburgh handedness inventory; Oldfield, 1971) native speakers of English, participated for

payment. All participants had normal hearing and vision, and no history of language impairment. All participants gave informed consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects (COUHES).

Design, materials and procedure.

Each participant completed a language localizer task (Fedorenko et al., 2010) and the critical gesture-processing task. In addition, participants performed a demanding spatial working memory (WM) task, which was used to define a set of control regions: the regions of the domain-general fronto-parietal multiple demand (MD) network (Duncan, 2010, 2013). These regions have been argued to support flexible cognition and goal-directed behaviors (e.g., Duncan & Owen, 2000; Fedorenko, Duncan, & Kanwisher, 2013). We included these regions because the MD network has been implicated in processing complex sequential stimuli, including biological actions (e.g., Culham & Valyear, 2006; Gallivan & Culham, 2015; Pritchett et al., 2018). In fact, some have labeled parts of this network the 'action observation network (AON)' (e.g., Biagi et al., 2015; Caspers, Zilles, Laird, & Eickhoff, 2010). Furthermore, some prior fMRI studies on gesture processing reported responses in what-appear-to-be parts of the MD network (e.g., Dick et al., 2009; Green et al., 2009; Willems et al., 2007). We therefore wanted to examine the responses of these regions to gesture processing. Some participants completed one or two additional tasks for unrelated studies. The entire scanning session lasted approximately 2 hours.

Language localizer: Participants passively read sentences and lists of pronounceable nonwords in a blocked design. The *Sentences* > *Nonwords* contrast targets brain regions sensitive to high-level linguistic processing (Fedorenko et al., 2010, 2011). Each trial started with 100 ms pre-trial fixation, followed by a 12-word-long sentence or a list of 12 nonwords presented on the screen one word/nonword at a time at the rate of 450 ms per word/nonword. Then, a line drawing of a hand pressing a button appeared for 400 ms, and participants were instructed to press a button whenever they saw this icon. Finally, a blank screen was shown for 100 ms, for a total trial duration of 6 seconds. The simple button-pressing task was included to help participants stay awake and focused. Each block consisted of 3 trials and lasted 18 s. Each run consisted of 16 experimental blocks (8 per condition), and five fixation blocks (14 s in duration each), for a total duration of 358 s (5 min 58 s). Each participant performed two runs. Condition order was counterbalanced across runs.

MD localizer: Participants performed a spatial working memory task that we have previously found to activate the MD system broadly and robustly (Fedorenko et al., 2013; Blank, Kanwisher, & Fedorenko, 2014). Subjects had to keep track of four (easy condition) or eight (hard condition) locations in a 3 × 4 grid (Fedorenko et al., 2011). In both conditions, subjects performed a two-alternative, forced-choice task at the end of each trial to indicate the set of locations that they just saw. The *Hard* > *Easy* contrast targets brain regions engaged in cognitively demanding tasks. Fedorenko et al. (2013; see also Hugdahl, Raichle, Mitra, & Specht, 2015) have shown that the regions activated by this task are also activated by a wide range of tasks that contrast a difficult vs. an easier condition. Each trial lasted 8 seconds (see Fedorenko et al., 2011, for details). Each block consisted of 4 trials and

lasted 32 s. Each run consisted of 12 experimental blocks (6 per condition), and 4 fixation blocks (16 s in duration each), for a total duration of 448 s (7 min 28 s). Each participant performed one or two runs. Condition order was counterbalanced across runs when participants performed two runs.

Critical gesture task: Participants listened to and/or watched excerpts from fairy tales narrated by eight different "actors" in a blocked design. The materials were created in five stages, as described next. All the materials are available at https://osf.io/zjpgf/.

Stage 1: Video recording: Eight volunteer undergraduate students (five females) at Princeton University were video-recorded while they narrated five fairy tales (The Queen Bee, The Leap Frog, Little Red Riding Hood, The Aged Mother, and The Princess and the Pea). The students were reminded of the plots of the fairy tales prior to the recording and were encouraged to use gestures. Each student was also recorded while engaging in a variety of non-gesture, mostly grooming, actions (e.g., scratching, fixing one's hair, stretching, putting lotion on one's hands, playing with a pencil, etc.), also known as self-adaptor movements (Ekman & Friesen, 1969; Dick et al., 2009). While engaging in these actions, the students were conversing with the cameraman so that they were producing speech during both types of movements. Narrations were not pre-rehearsed so that they would look and sound naturalistic.

Stage 2: Selecting the gesture and non-gesture (grooming) clips: One hundred and fifty clips were selected from the gesture recordings (18-19 per actor, 20-33 per fairy tale). These clips had to (a) be between 5.5 s and 7 s in duration, (b) include gesture production (see below for more details on the types of gestures produced), and (c) include a self-contained linguistic message (typically, a sentence or two; transcriptions available at https://osf.io/zjpgf/). For each of 150 gesture clips, a 6 s clip from among the grooming recordings of the same actor was chosen.

Stage 3: Editing the audio for the non-gesture clips: The audio signal was removed from the non-gesture clips, and replaced by the audio signal from the corresponding gesture clip. Thus, the audio was identical across the two video+audio conditions. (If the audio was shorter than 6 s, it started slightly after the grooming video began and/or ended slightly before the video ended; if the audio was longer than 6 s, it was cut off, with a fade-out, at 6 s.).

Stage 4: Editing the face out: Each video was edited so that a beige-colored rectangle covered the participant's face (Figure 2a). This was done for two reasons: First, we wanted to use the same audio signal across the gesture and grooming conditions to minimize potential between-condition differences not related to the critical movement-type manipulation, and this would not have been possible without the face masking (because there would have been mismatches between the articulators and the audio signal in the grooming condition). And second, we wanted participants to focus on processing the hand and arm movements, and if the faces were not covered, they would have attracted the most attention (e.g., Sato & Kawahara, 2015).

Stage 5: Creating the video-only and audio-only conditions: The video-only conditions were created by removing the audio signal, and the speech-only condition was created by just using the audio from the gesture clips. The audio signals were intensity-normalized using the Audacity recording and editing software.

Each of 150 items therefore had five versions, corresponding to the five experimental conditions (SpeechOnly, SilentGesture, SilentGrooming, SpeechGesture, SpeechGrooming), for a total of 750 trials. These trials were divided into five experimental lists (150 trials each, 30 trials per condition) following a Latin Square design so that each list contained only one version of an item. Any given participant saw the materials from just one experimental list.

Each trial lasted 7 s (clips shorter than 7 s were "padded" with silence and a blank screen at the end). Each experimental block consisted of 3 trials and lasted 21 s. Trials were grouped into blocks in a way that minimized overlap in actors and fairy tales within each block. Each run consisted of 10 experimental blocks (2 per condition) and 3 fixation blocks each lasting 16 s, for a total run duration of 258 s. Each participant performed 5 runs. (Due to a technical error, four participants saw only a subset of the items from the target list: the same 30 trials [6 per condition] were presented in each of five runs. However, given that the results of the analyses with vs. without these participants were similar, we included the data from these participants.) Condition order was counterbalanced across runs and participants.

Characterization of the gesture and grooming materials: To better characterize the experimental materials, we analyzed the amount of motion, and the number and type of movements produced, and conducted four behavioral studies, as described below. All the data are available at https://osf.io/zjpgf/.

Amount of motion: To test whether the gesture and grooming videos were similar in motion energy, we used the Computer Vision System Toolbox for MatLab to estimate the distribution of the apparent velocities of objects in each video. In particular, for each pixel in each frame, we identified the amount of displacement between frames (pixel velocity). Next, we summed the pixel velocities per frame to get an overall motion energy for each frame. Finally, we calculated an average motion energy for each video. The ranges and distributions of motion energy values were similar between conditions (Figure 2B), with a few grooming videos with very high motion energy and more gesture videos with low motion energy leading to a significant difference (M(gesture) = 16,058, SD = 10,345; M(grooming) = 19,161, SD =11,443; $t(149) = 2.46$, $p = .02$). (To briefly foreshadow the results, this difference in motion energy could potentially explain higher neural responses to the grooming than the gesture videos in several fROIs in the MD network. However, the language fROIs did not show a reliable difference between the grooming and gesture videos, suggesting that this difference in motion energy does not affect the key results.)

Number of movements: To test whether the gesture and grooming videos were similar in terms of the number of individual movements, each video was segmented into movement phrases, following the guidelines described in Kendon (1980) and Kita, Van Gijn, and Van der Hulst (1998). A movement phrase was defined as consisting of the following phases: preparation (the movement of the hand as it readies itself for the gestural stroke), stroke (the

most effortful and meaningful phase of a movement), and retraction (the movement of the hand as it returns to a resting position or to a position required by a preparation phase of a subsequent movement). In some instances, strokes are preceded or followed by holds that are considered to be part of the same movement phrase. Two raters (O.J. and Z.L.) independently segmented gesture and grooming videos into movement phrases, and the inter-rater agreement for the number of movements was computed using Cohen's kappa coefficient (k). Agreement was high for both gesture ($k = .70$, $p < .001$) and grooming videos ($k = .66$, $p < .001$; values of k greater than .60 are considered acceptable; Fleiss & Cohen, 1973). The two raters then went over the videos where the movement counts diverged and reached agreement. The gesture and grooming videos were well-matched on the number of distinct movements (M(gesture) = 3.6, $SD = 1.11$; M(grooming) = 3.7, $SD = 1.75$; $t(298) = 0.40$, $p = 0.69$; Figure 2C).

Type of movements: The individual gesture movements (total count: 539) were categorized by Z.L., with input from S.G.-M., a world expert on gestures, into five categories (e.g., Abner et al., 2015): 1) beats (gestures that accentuate the topic without directly referring to it), 2) deictic gestures (gestures that point at an object being discussed), 3) iconic gestures (gestures that depict object attributes, spatial locations, and/or actions), 4) metaphoric gestures (gestures that depict abstract imagery), and 5) emblems (conventionalized gestures that convey meaning independent of speech). The distribution of gesture types is shown in Figure 2D. Iconic gestures were most common (~37%), followed by beats (~24%), metaphoric gestures (~20%), emblems (~15.5%), and deictic gestures (~3.5%). We also asked how often iconic and metaphoric gestures carried information that was supplementary to speech (e.g., pointing to one's stomach while mentioning that the character got sick; this gesture specifies the type of ailment) vs. simply congruent with it (e.g., pointing to one's head while mentioning that the character was smart). The vast majority of iconic and metaphoric gestures (91%) were of the latter type.

Behavioral studies: Participants: Across four studies, 369 participants were recruited via Amazon.com's online survey platform, Mechanical Turk, for payment. Seventy-five participants were excluded because (i) their native language was not English, (ii) they failed to complete the study, or (iii) they were outliers based on their performance as identified by the median absolute deviation test with a modified z-score threshold of 2.5. This left 294 participants for analysis (52 for Study 1, 92 for Study 2, 52 for Study 3, and 49 for Study 4).

Behavioral study 1 (n=52): Gesture vs. grooming differentiation: The goal of Study 1 was to test whether participants could distinguish between gesture and grooming videos. In version 1a, the videos were accompanied by the audio signal, and in version 1b, silent videos were used. In each version, the 300 videos were divided into two experimental lists, each containing 150 videos (75 gestures and 75 grooming), and any given participant saw one list. Participants were given the following instructions: "People often move their hands (gesture) when they talk. Gestures can be used to help express ideas or to highlight/emphasize something in the narrative. Other times, people may move their hands, but not to gesture. For example, someone may adjust their glasses, or fix their hair, or blow their nose. We want to see how well people can distinguish gesture movements from these non-gesture

("grooming") movements. You will watch short video clips ("video clips with an accompanying narrative" in Study 1a, or "silent video clips" in Study 1b) of people's hand movements, and your task is to decide — for each clip — whether the movements are gestures or grooming movements. Please, use a scale from 1 (confident the movements are gestures) to 5 (confident the movements are grooming movements)."

Participants could easily distinguish between gestures and grooming movements either with audio (M(gestures) = 1.09; M(grooming) = 4.60; $t(298) = -89.31$, $p < 0.001$), or without audio (M(gestures) = 1.12; M(grooming) = 4.37; $t(298) = -68.36$, $p < 0.001$). If we treat responses 1 and 2 as correctly identifying a video as a gesture video, and responses 4 and 5 as correctly identifying a video as a grooming video, participants, on average, identified 99% of the gesture videos and 88% of the grooming videos correctly, with audio, and 98% of the gesture videos and 84% of the grooming videos correctly, without audio (Figure 2E). In summary, the gesture vs. grooming manipulation was robust, and participants could clearly discriminate between the two conditions.

Behavioral study 2 (n=92): Familiarity of the gesture vs. grooming movements: The goal of Study 2 was to assess the familiarity of the movements used in the gesture vs. grooming videos. Here, we used the videos of the individual movements following the splitting of the videos into individual movements described in the section on Number of Movements, for a total of 539 movements from the gesture videos, and 549 movements from the grooming videos. The 1,088 video clips were divided into 8 lists (lists 1-7 each contained 70 gesture movements and 70 grooming movements, and list 8 contained 49 gesture movements and 59 grooming movements). Any given participant saw one list. Participants were given the following instructions: "People often move their hands when they talk. Some movements are used a lot and may seem very familiar; other movements may be used only occasionally and may seem less familiar. You will watch clips of people's hand movements, and your task is to rate each movement on how familiar it is. Please, use a scale from 1 (very unfamiliar, I have almost never seen this movement) to 5 (very familiar, I have seen this movement often)."

Participants rated both gesture and grooming movements as quite familiar (above 3 on the scale above), with the grooming movements rated as more familiar ($M = 3.76$, $SD = 0.67$) than gesture movements ($M = 3.14$, $SD = 0.74$; $t(1086) = 14.67$, $p < 0.001$; Figure 2F).

Behavioral study 3 (n=52): Communicativeness of the gesture vs. grooming movements: The goal of Study 3 was to assess the communicative nature of the movements used in the gesture vs. grooming videos. As in Study 2, we used the videos of the individual movements. The 1,088 video clips were divided into lists in the same way as in Study 2. Any given participant saw one list. Participants were given the following instructions: "People often move their hands when they talk. Sometimes, those movements can be used to help express ideas or to highlight/emphasize something in the narrative. Other times, the movements don't have anything to do with what the person is saying (for example, when a person is shuffling cards while talking). You will watch clips of people's hand movements, and your task is to decide — for each clip — whether the movement was produced to aid with the narration (you will not hear the audio, just see the movement). Please use a scale

from 1 (confident the movement was produced to aid with the narration) to 5 (confident the movement had nothing to do with the narration).”

Participants rated gestures as much more communicative ($M = 1.60$, $SD = 0.57$) than grooming movements ($M = 4.05$, $SD = 0.60$; $t(1086) = 69.17$, $p < 0.001$; Figure 2G). Together with Study 1, the results of this study demonstrate that the two conditions are clearly distinguishable and that, as expected, gestures are rated as more communicative than grooming movements.

Behavioral study 4 (n=49): Interpretability of the gesture movements.: In Study 3, we found that participants judge gestures to be more communicative. The goal of Study 4 was to assess the actual information contained in the gesture movements in the absence of speech. As in Studies 2 and 3, we used the videos of the individual movements, but here, we only included the movements from the gesture condition. The 539 video clips were divided into 4 lists (lists 1-4 each contained 140 movements, and list 4 contained 119 movements). Any given participant saw one list. Participants were given the following instructions: “People often move their hands (gesture) when they talk. Sometimes, those movements can be interpreted on their own (without the accompanying speech); other times, it’s hard to figure out what the gestures mean without hearing the speech. You will watch clips of people’s hand movements produced while they talked about something, and your task — for each clip — is to try to understand what the gesture means and provide a word, a couple of words, or a phrase to describe it. For example, if you see someone holding out three fingers, you could write “three”; or if you see someone spreading their arms wide, you could write “big” or “caught a big fish”; or if you see someone shaking their finger, you could write “I told you so” or “You are naughty”. In some cases, it may be difficult to interpret the gesture, so just make your best guess.”

To estimate the information contained in the gesture movements, we examined the amount of lexical overlap in the participants’ productions. To quantify the overlap, we used entropy, an information-theoretic measure (e.g., Cover & Thomas, 1991) that captures how surprising different elements are in a set. Before computing entropy, the elicited productions were preprocessed using Python’s Natural Language Toolkit: all the words were lower-cased and lemmatized (e.g., “big” and “bigger” would be treated as the same), punctuation marks were removed, and repeated words within a production were removed (e.g., the phrase “knock, knock, knock” would be reduced to “knock”). We then used Term Frequency-Inverse Document Frequency (TF-IDF) to minimize the influence of highly frequent, but not very informative, function words, like determiners and pronouns, in the entropy calculations. In particular, we used the formula TF-IDF = $tf_{i,j} * log(N/df_i)$, where $tf_{i,j}$ is the number of occurrences of word i across all the productions (document j), df_i is the number of productions that contain word i (note that, in our case, the latter two terms were identical because repeated words within a production were removed), and N is the total number of productions (7,007 in our case). Using a TF-IDF score threshold of 0.048 we filtered out the most common function words (“it”/“its”, “I”, “be”/“am”, “you”, “and”, “a”, “the”, “this”, “that”).

We then computed entropy, using the following formula:

$$Entropy\ H(X) = -k\sum_{i=1}^{n} p_i log_2(p_i)$$

where p_i is the empirical probability that word i was produced to describe the gesture movement, n is the number of unique words produced to describe the movement across all productions for that movement (on average, 3.8, $SD = 2.43$), and k is a normalizing factor (k=1 for entropy based only on word probabilities, and k = 1/(average production length for a movement) to account for differences in production length across participants).

The average entropy across gesture movements was 3.97 bits ($SD = 0.62$) and the average normalized entropy was 1.85 bits ($SD = 0.41$; Figure 2H). These values suggest that, on average, only 3.97 yes/no questions would be required to guess the word used to describe the movement from among all the words in the productions, or 1.85 yes/no questions to guess the word while also accounting for phrase length. These results demonstrate that participants' responses contained a lot of lexico-semantic overlap. For comparison, consider two extreme cases: a) a case in which every participant produced the same word to describe a movement, and b) a case in which every participant produced a unique response to describe a movement. In the former, entropy is 0. And in the latter case, the average entropy is significantly higher than what was observed in our study ($M = 4.57$, $SD = 0.28$; $z = 20.4$, $p < .0001$). As expected, the amount of lexical overlap further varied across gesture types: the entropy was lowest (indicating highest lexical overlap) for emblems ($M = 3.59$, $SD = 1.03$), followed by iconic gestures ($M = 4.00$, $SD = 0.58$), metaphoric gestures ($M = 4.10$, $SD = 0.36$), deictic gestures ($M = 4.20$, $SD = 0.39$), and, finally, beats ($M = 4.28$, $SD = 0.29$).

fMRI data acquisition and preprocessing.

Structural and functional data were collected on the whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 176 sagittal slices with 1 mm isotropic voxels (TR = 2,530 ms, TE = 3.48 ms). Functional, blood oxygenation level dependent (BOLD) data were acquired using an EPI sequence (with a 90° flip angle and using GRAPPA with an acceleration factor of 2), with the following acquisition parameters: thirty-one 4 mm thick near-axial slices, acquired in an interleaved order with a 10% distance factor; 2.1 mm x 2.1 mm in-plane resolution; field of view of 200 mm in the phase encoding anterior to posterior (A > P) direction; matrix size of 96 mm x 96 mm; TR of 2,000 ms; TE of 30 ms. Prospective acquisition correction (Thesen, Heid, Mueller, & Schad, 2000) was used to adjust the positions of the gradients based on the participant's motion one TR back. The first 10 s of each run were excluded to allow for steady-state magnetization.

MRI data were analyzed using SPM5 and custom MATLAB scripts. (Note that SPM was only used for preprocessing and basic first-level modeling – aspects that have not changed much in later versions; we used an older version of SPM because data for this study are used across numerous other projects spanning many years and hundreds of participants, and we wanted to keep the SPM version the same across all the participants.) Each participant's data

were motion corrected and then normalized into a common brain space (the Montreal Neurological Institute, MNI, Brain Template) using 16 nonlinear iterations with 7×9×7 basis functions and resampled into 2 mm isotropic voxels. The data were then smoothed with a 4 mm Gaussian filter and high-pass filtered (at 200 s). The localizer tasks' and the critical task's effects were estimated using a General Linear Model (GLM) in which each experimental condition was modeled with a boxcar function convolved with the canonical hemodynamic response function (HRF).

Definition of group-constrained, subject-specific fROIs.

In contrast to prior fMRI studies of co-speech gesture processing, which relied on traditional group-averaging analyses, we analyzed the percent BOLD signal change within regions of interest that were defined functionally for each individual participant. For each participant, functional regions of interest (fROIs) were defined using the Group-constrained Subject-Specific (GSS) analysis method (Fedorenko et al., 2010; Julian, Fedorenko, Webster, & Kanwisher, 2012). In this method, a set of parcels or "search spaces" (i.e., brain areas within which most individuals in prior studies showed activity for the localizer contrast) is combined with each individual participant's activation map for the same contrast. The parcels are large by design in order to account for inter-individual variability in the precise locations of functional regions.

To define the language fROIs, we used five parcels derived from a group-level representation of data for the *Sentences* > *Nonwords* contrast in 220 participants (Figure 3A). These parcels included three regions in the left frontal cortex: two located in the inferior frontal gyrus (LIFG and LIFGorb), and one located in the middle frontal gyrus (LMFG); and two regions in the left temporal and parietal cortices spanning the entire extent of the lateral temporal lobe (LAntTemp, LPostTemp). These parcels are similar to the parcels reported originally in Fedorenko et al. (2010) based on a set of 25 participants, except that the two anterior temporal parcels (LAntTemp, and LMidAntTemp) are grouped together, and two posterior temporal parcels (LMidPostTemp and LPostTemp) are grouped together. Further, we did not include the LAngG parcel because this fROI has been shown to consistently pattern differently from the rest of the language network across diverse measures (e.g., Blank et al., 2014; Chai, Mattar, Blank, Fedorenko, & Bassett, 2016; Mineroff, Blank, Mahowald, & Fedorenko, 2018; Pritchett et al., 2018), including responding more strongly to visual meaningful stimuli than to sentences (Amit, Hoeflin, Hamzah, & Fedorenko, 2017), which suggests that it is not a language region. The language as well as the MD (see below) parcels are available for download from the Fedorenko lab website.

To define the MD fROIs, following Fedorenko et al. (2013) and Blank et al. (2014), we used eighteen anatomical regions across the two hemispheres (Tzourio-Mazoyer et al., 2002; Figure 3B) previously implicated in MD activity: opercular IFG (LIFGop & RIFGop), MFG (LMFG & RMFG), orbital MFG (LMFGorb & RMFGorb), insular cortex (LInsula & RInsula), precentral gyrus (LPrecG & RPrecG), supplementary and presupplementary motor areas (LSMA & RSMA), inferior parietal cortex (LParInf & RParInf), superior parietal cortex (LParSup & RParSup), and anterior cingulate cortex (LACC & RACC).

Within each language and MD parcel, we selected the top 10% of most localizer-responsive voxels based on the *t* values for the relevant contrast (*Sentences > Nonwords*, and *Hard > Easy* spatial WM, respectively; see e.g., Figure 1 in Blank et al., 2014, for sample fROIs). This approach ensures that a fROI can be defined in every participant, and that the fROI sizes are identical across participants.

To estimate the responses of fROIs to the conditions of the corresponding localizers – important for ensuring that these regions behave as expected – we used an across-run cross-validation procedure (Nieto-Castañon & Fedorenko, 2012). In this procedure, (i) the first run is used to define the fROIs, and the second run to estimate the response; then, (ii) the second run is used to define the fROIs and the first run to estimate the responses; and finally, (iii) the estimates are averaged across the two runs to derive a single value per participant per region. To estimate the responses of fROIs to the conditions of the critical task, data from all the localizer runs were used to define the fROIs. The extracted response magnitude values (in percent BOLD signal change) were used for second-level analyses.

Analyses.

To address our primary research question – whether language regions respond to gestures, i.e., non-linguistic communicatively-relevant signals (Figure 1A) – we first examined the responses of the language fROIs to the two silent video conditions. For each region, we performed two paired-samples *t*-tests, one comparing the critical *SilentGesture* condition to the *SpeechOnly* condition, and the other comparing the *SilentGesture* condition to the control *SilentGrooming* condition.

To further evaluate the additional hypotheses outlined above (namely, that the language regions respond to the amount of communicatively-relevant information, Figure 1C; or to speech-gesture integration demands, Figure 1D), we performed two paired-samples *t*-tests, one comparing the *SpeechGesture* condition to the *SpeechOnly* condition, and the other comparing the *SpeechGesture* condition to the *SpeechGrooming* condition.

The same *t*-tests were performed on each of the MD regions.

To control the false-positive rate when tests on multiple regions were performed, the False Discovery Rate (FDR) correction was applied to control for the number of regions in each network, i.e., five in the language network, and eighteen in the MD network (Benjamini & Yekutieli, 2001).

The individual-subjects fROI approach is characterized by high statistical power (e.g., Saxe, Brett, & Kanwisher, 2006; Nieto-Castañon & Fedorenko, 2012). However, it has been criticized with respect to two potential limitations (e.g., Friston, Rotshtein, Geng, Sterzer, & Henson, 2006). First, it may obscure potential functional heterogeneity within fROIs. For example, if a fROI contains only a small proportion of voxels that respond to communicatively-relevant signals, like gestures, we may miss these responses by averaging across all the voxels within the fROI. And second, by focusing on only a subset of the brain (i.e., on the regions that show a particular functional signature), we may miss important regions in other parts of the brain. Although the second criticism is not relevant given the

scope of the primary research question in the current study, which concerns the response to gesture of the language processing mechanisms whose neural substrates have been well characterized previously, we performed several additional analyses to address these potential limitations, which are described in the Results section.

For ease of comparison with prior studies, we further performed a traditional whole-brain random-effects group analysis for the following contrasts: *Gesture > Grooming*, Grooming > Gesture, *Gesture > Fixation, and Grooming > Fixation.* The results are reported at https://osf.io/zjpgf/.

Results

Validating the language and MD fROIs

As expected and replicating previous work (e.g., Fedorenko et al., 2010, 2011; Mahowald & Fedorenko, 2016), the *Sentences > Nonwords* effect was reliable in each of five language fROIs (ts(16) ≥ 5.45, ps < 0.001). Similarly, the *Hard > Easy* spatial WM effect was reliable in each of eighteen MD fROIs (ts(16) ≥ 2.93, ps < 0.004). Thus both sets of fROIs show the expected functional signatures, and we can proceed to examine their responses to the conditions of the critical experiment.

Responses of the language fROIs to the critical gesture experiment

As can be clearly seen from Figure 3B, the language regions showed a robust response to the *SpeechOnly* condition, replicating Scott et al. (2016), They also responded strongly to the two video conditions with speech (*SpeechGesture* and *SpeechGrooming).* However, the response to the two silent video conditions was low, barely above the fixation baseline. The t-tests (Table 1) revealed that the *SpeechOnly* condition elicited a reliably stronger response than the *SilentGesture* condition (ts(16) ≥|−2.67|, ps < 0.02), but the *SilentGesture* condition did not differ from the *SilentGrooming* condition (ts(16) ≤|1.71|, ps > 0.11). Furthermore, although adding gestures to the speech signal led to a reliable increase in response magnitude in all language fROIs (ts(16) ≥|3.72|, ps > 0.002), the *SpeechGesture* condition did not differ from the *SpeechGrooming* condition in any of the regions (ts(16) ≤|1.60|, ps > 0.13).

A visual examination of the MD fROIs' profiles (Figure 3B) makes it clear that their response to the gesture task differs from that of the language fROIs. In particular, the responses to all the video conditions were quite strong in many MD regions, and at least some MD regions responded more strongly to the *Grooming* than the *Gesture* conditions, in both silent videos and videos with speech, although these differences did not survive the multiple comparisons correction. Furthermore, unlike the language regions, which responded more strongly to the conditions with the linguistic signal, the MD regions showed the opposite pattern numerically, with stronger responses to the silent video conditions.

Potential heterogeneity of language fROIs?

To address the issue of potential heterogeneity within language fROIs, we performed two additional analyses. For both of these, we used the *Gesture > Grooming* contrast, collapsing

across the silent and with-speech conditions. This was done to maximize our power to detect language-gesture overlap. In the first analysis, we examined voxel-level overlap between the language localizer contrast (*Sentences > Nonwords*) and the *Gesture > Grooming* contrast within the language parcels. To do so, we counted the number of voxels that show a significant response to each contrast (at the whole-brain threshold level of $p < 0.001$) and then quantified the overlap between the two sets using the Jaccard index (i.e., the number of overlapping voxels divided by the total number of voxels for the two contrasts, multiplied by 100). For the *Gesture > Grooming* contrast, we additionally examined voxels that show a significant response at more liberal thresholds ($p < 0.01$ and < 0.05), to give the language-gesture overlap the best chance to reveal itself.

The volumes of the language fROIs were significantly larger than those of the gesture-responsive regions, even when the gesture-responsive regions were defined using the most liberal threshold ($p < 0.05$), ($t(16) = 3.34$, $p = 0.001$; Figure 4A). On average, 31% of voxels within the language parcels showed significant *Sentences > Nonwords* responses (at the whole-brain threshold level of $p < 0.001$, uncorrected), whereas only 2% of voxels responded to the *Gesture > Grooming* contrast at the same threshold level. Even at more liberal thresholds, the volumes of the gesture processing regions were relatively small (7% at $p < 0.01$; 14% at $p < 0.05$). Across the language network, voxel-level overlap between the language contrast and the gesture contrast was only 3% ($SD = 4\%$) at the whole-brain threshold of $p < 0.001$ for each contrast. Within most parcels, the majority of participants (range 10-17 of 17) showed no overlap between language-responsive and gesture-responsive voxels. The only parcel where the majority of participants (10 of 17) showed *some* overlap was the LPostTemp parcel, but the amount of overlap was still low (*M Jaccard Index* across participants = 4%, $SD = 5\%$, whole-brain threshold level of $p < 0.001$, uncorrected; Figure 4B). And critically, the *magnitude* of response to the gesture condition with no linguistic signal was low, as shown in the main analysis (Figure 3A).

In a second analysis, we tested whether the strength of the response to language processing is related to the strength of the response to gesture processing across the voxels in each region of interest. In other words, do the same voxels that show stronger responses to language also show stronger responses to gesture processing? Similar activation landscapes for language processing and gesture processing should yield high correlations. To perform this analysis, we extracted – in each participant – the contrast values in each voxel of each language parcel for (a) the *Sentences > Nonwords* contrast of the language localizer, (b) the *Gesture > Grooming* contrast of the critical experiment, and, for comparison, (c) the *Speech > NoSpeech* contrast of the critical experiment (which is expected to be similar in its activation landscape to the language localizer contrast; e.g., Scott et al., 2016). These contrast values were extracted separately from the odd and even runs in order to estimate the replicability of activation patterns within each contrast, which provides the upper bound for the potential similarity between contrasts. A series of correlations were then computed on these extracted sets of contrast values in each participant and parcel: (i) the *Sentences > Nonwords* odd vs. even runs, (ii) the *Gesture > Grooming* odd vs. even runs, (iii) the *Speech > NoSpeech* odd vs. even runs, and, critically, (iv) the *Sentences > Nonwords* odd runs vs. the *Gesture > Grooming* even runs, (v) the *Sentences > Nonwords* even runs vs. the *Gesture > Grooming* odd runs, (vi) the *Sentences > Nonwords* odd runs vs. the *Speech > NoSpeech*

even runs, and (vii) the *Sentences > Nonwords* even runs vs. the *Speech > NoSpeech* odd runs. The correlation values were Fisher-transformed (Fisher, 1921), to improve normality (Silver & Dunlap, 1987). For each participant and parcel, we averaged the values obtained in (iv) and (v) above (*Sentences > Nonwords* / *Gesture > Grooming* correlation), as well as in (vi) and (vii) (*Sentences > Nonwords* / *Speech > NoSpeech* correlation), to derive a single correlation value for the between-contrast comparisons. Finally, these between-contrast correlations were tested a) against the null baseline, and b) directly against each other (i.e., comparing *Sentences > Nonwords* / *Gesture > Grooming* correlation values vs. *Sentences > Nonwords* / *Speech > NoSpeech* correlation values) via simple *t*-tests.

The activation patterns within the language parcels were highly replicable for the *Sentences > Nonwords* contrast of the language localizer task (*M*(*r, Fisher transform*) = 1.16, *range*(*r, Fisher transform*) = 1.01-1.22, *p*s < 0.001; see also Mahowald & Fedorenko, 2016) and for the *Speech > NoSpeech* contrast of the critical experiment (*M*(*r, Fisher transform*) = 1.15, *range*(*r, Fisher transform*) = 0.72-1.78, *p*s < 0.001). Further, the activation patterns for the *Sentences > Nonwords* vs. *Speech > NoSpeech* contrasts were also significantly correlated with each other in all language parcels (*M*(*r, Fisher transform*) = 0.45, *range*(*r, Fisher transform*) = 0.29-0.59, *p*s < 0.001), in line with Scott et al. (2016). However, the activation patterns for the *Gesture > Grooming* contrast of the critical experiment were not replicable across runs (*M*(*r, Fisher transform*) = 0.10, *range*(*r, Fisher transform*) = 0.02-0.17; all *p*s *n.s.*), and, correspondingly, not correlated with the patterns for the *Sentences > Nonwords* contrast (*M*(*r, Fisher transform*) = 0.15, *range*(*r, Fisher transform*) = 0.04-0.22, all *p*s *n.s.).* The latter correlations were significantly lower than the *Sentences > Nonwords* vs. *Speech > NoSpeech* correlations (*t*s > 1.99, *p*s < 0.03).

Regions of language-gesture overlap elsewhere in the brain?

To address the issue of potentially missing important regions of overlap between language and gesture processing outside the boundaries of our language parcels, we performed a whole-brain Group-constrained Subject-Specific (GSS) analysis, which searches for spatially consistent regions of activation across participants, while taking into account inter-individual differences in the precise locations of functional activations (Fedorenko et al., 2010). For this analysis, in each participant individually, we identified voxels – across the brain – that respond to both the language localizer contrast (*Sentences > Nonwords*) and the *Gesture > Grooming* contrast, each at the liberal whole-brain threshold level of $p < 0.05$ uncorrected, to give the language/gesture overlap hypothesis the strongest chance. We then overlaid these individual activation maps on top of one another and searched for regions that are spatially consistent across participants (see Fedorenko et al., 2010, for details of the procedure; see Fedorenko et al., 2015, for previous uses of the whole-brain GSS analysis). Meaningful regions were defined as regions that (a) are present (i.e., have at least one above-threshold voxel within their boundaries) in at least 80% of participants, (b) show replicable *Sentences > Nonwords* and *Gesture > Grooming* effects in a left-out portion of the data (tested with across-runs cross-validation, as described above), and (c) respond to each of *Sentences* and *Gesture* conditions reliably above the fixation baseline.

Even at the very liberal whole-brain threshold of $p < 0.05$, we detected no spatially systematic regions in which activations for linguistic and gesture processing tasks overlapped. This result suggests that language and gestures are processed by distinct sets of brain areas, in line with the main analysis (Figure 3A).

Searching for gesture- and grooming-responsive regions across the brain

Finally, to ensure that our gesture vs. grooming manipulation was effective, we used a whole-brain GSS analysis described above to search for gesture-responsive regions across the brain, regardless of whether these regions overlap with the language regions; we also searched for grooming-responsive regions across the brain. We first tried this analysis with individual activation maps for the *Gesture* > *Grooming* and *Grooming* > *Gesture* contrasts thresholded at $p < 0.001$. Given that no meaningful regions emerged for the *Gesture* > *Grooming* contrast (by a similar definition as the one above, except for lowering the first criterion to 70% of participants, and disregarding the responses to the conditions of the language localizer), we then thresholded individual *Gesture* > *Grooming* maps liberally at $p < 0.05$, and recovered a number of gesture-responsive regions. (Note that although a liberal threshold is used for the individual maps, the observed regions' responses are cross-validated across runs, which ensures that the responses are internally replicable.)

In the whole-brain GSS analysis, no spatially systematic regions emerged for the *Gesture* > *Grooming* contrast when individual activation maps were thresholded at $p < 0.001$. The analysis with liberally thresholded individual maps ($p < 0.05$) discovered one region that was present in more than 80% of participants, and four additional regions that were present in more than 70% but fewer than 80% of participants (Figure 5A; Table 2). Two regions were located bilaterally in the posterior temporal cortex, one in the left superior frontal gyrus, one in the medial frontal cortex, and one in the right cerebellum. The *Gesture* > *Grooming* effect was replicable, as assessed with across-runs cross-validation, in each of the five fROIs (ts > 2.30, ps < 0.02; Figure 5B).

For the Grooming > Gesture contrast (with the individual activation maps thresholded at $p < 0.001$), three regions emerged that were present in more than 70% but fewer than 80% of participants (Figure 5C; Table 3). Two regions were located bilaterally in the superior parietal cortex and one in the right inferior parietal cortex. The Grooming > Gesture effect was replicable, as assessed with across-runs cross-validation, in each of the three fROIs (ts > 2.57, ps < .02; Figure 5D). These results are in line with the MD fROIs located in similar anatomical locations exhibiting stronger responses to grooming than gesture, as discussed above.

Discussion

We asked whether the processing of co-speech gestures – the movements of hands and arms that often accompany spoken utterances – engages the brain regions that support language comprehension. A number of prior studies have reported responses to gesture manipulations within the same macroanatomical areas as the ones implicated in language processing (e.g., Dick et al., 2009; Holle et al., 2008; Holler et al., 2015; Hubbard et al., 2009; Weisberg et al., 2016) and interpreted those responses as evidence of shared processing mechanisms.

However, as discussed in the introduction, prior studies have suffered from one or more of the following limitations: they (a) have not directly compared activations for gesture and language processing in the same study and relied on the fallacious reverse inference (Poldrack, 2006) for interpretation, (b) relied on traditional group analyses, which are bound to overestimate overlap (e.g., Nieto-Castañon & Fedorenko, 2012), (c) failed to directly compare the magnitudes of response (e.g., Chen et al., 2017), and (d) focused on gestures that may have activated the corresponding linguistic representations (e.g., "emblems"). To circumvent these limitations, we examined responses to gesture processing in language regions defined functionally in individual participants (e.g., Fedorenko et al., 2010), including directly comparing effect sizes, and covering a broad range of spontaneously generated co-speech gestures.

In line with prior work (e.g., Fedorenko et al., 2010; Scott et al., 2016), we found robust responses to conditions where a linguistic (in this case, auditory) signal was present, in all language regions[1]. Critically, however, we found no evidence that any of the language regions respond more strongly to gesture processing in the absence of speech than to processing grooming movements that can accompany speech but have no relation to the linguistic signal. Both of the silent video conditions (*SilentGesture* and *SilentGrooming*) produced a low response – at or slightly above the fixation baseline – in all language regions, suggesting that the presence of gestures alone is not sufficient to activate these areas. These results argue against the hypothesis whereby the language regions respond to any communicatively-relevant signal (Figure 1A).

Further, although some regions showed stronger responses to the condition where the speech signal was accompanied by gestures compared to the *SpeechOnly* condition, the *SpeechGesture* condition did not reliably differ from the *SpeechGrooming* condition. So, the higher responses to the *SpeechGesture* compared to the *SpeechOnly* condition are likely due to the fact that having video is more engaging than not having video. Moveover, the contributions of the linguistic signal vs. the gestures were highly uneven: whereas the magnitude of response to the *SpeechOnly* condition was 60-65% of the *SpeechGesture* condition, the magnitude of responses to the *SilentGesture* condition was only 6-33% of the *SpeechGesture* condition, suggesting that the presence of the linguistic signal, not the presence of gestures, is the primary driver of the response to the *SpeechGesture* condition. This pattern of results argues against the hypothesis whereby the language regions respond to the total amount of communicatively-relevant information (Figure 1C), the hypothesis whereby the language regions respond to the demands associated with the need to integrate communicatively-relevant information from different signals (Figure 1D; cf. Demir-Lira et al., 2018; Dick et al., 2009; 2012), and the hypothesis whereby the language regions respond to communicatively-relevant signals, but gestures only become communicatively-relevant in the presence of speech (Figure 1D). Instead, these results strongly support the hypothesis

[1]Note that the strong response to the *SpeechOnly* condition – a naturalistic auditory linguistic signal – in the language regions, including regions in the inferior frontal gyrus (see also Scott et al., 2016), is contra prior claims that such stimuli only produce responses in the temporal cortices (e.g., Friederici, Meyer, & von Cramon, 2000; Humphries, Binder, Medler, & Liebenthal, 2006; Rogalsky & Hickok, 2009). This result, once again, highlights the importance of the functional localization approach, which allows the detection of robust effects that are entirely missed by traditional group analyses plagued by high interindividual variability (e.g., Saxe, Brett, & Kanwisher, 2006; Nieto-Castañon & Fedorenko, 2012; Glezer & Riesenhuber, 2013)

that the language regions are selective for linguistic content (Figure 1B; e.g., Fedorenko et al., 2011).

To ensure that averaging responses across the voxels within the language fROIs did not obscure potential responses to gesture processing in a small subset of language-responsive voxels, we conducted two additional analyses. The results were consistent with the analysis of the language fROIs' response profiles. In particular, within most language parcels we found minimal or no overlap between language- and gesture-responsive voxels in the vast majority of participants. The only language parcel within which a substantial proportion of participants showed a small amount of overlap (~4% of voxels) was the LPostTemp parcel. Even so, the fine-grained patterns of activation for the *Gesture > Grooming* contrast (i) were not even reliable within participants across runs, suggesting that these responses may be dominated by noise (or possibly driven by a small subset of the trials), and (ii) were distinct from those for the language contrasts (the *Sentences > Nonwords* contrast in the language localizer, or the *Speech > NoSpeech* contrast in the critical experiment), as evidenced by low correlations within participants (i.e., there was no relationship between how strongly any given voxel responded to language processing vs. gesture processing). And as we saw in the main analysis, the magnitude of the response to gestures with no linguistic signal is close to the fixation baseline.

One thing to keep in mind is that we have here focused on (relatively) naturalistic co-speech gestures, which span diverse kinds of gestures, including beats, deictic gestures, iconic gestures, metaphoric gestures, and emblems (Figure 2D). The choice of naturalistic gesture productions with a-few-seconds-long clips precluded us from being able to separate neural responses to individual movements and therefore to examine potential differences among the different types of gestures. As discussed in the introduction, some researchers have argued that *particular kinds* of co-speech gestures may elicit responses in the language areas. Some of these claims have been made with respect to language-gesture integration: gestures that clarify the linguistic message have been argued to elicit a stronger response than conditions where such gestures are not needed to understand the message (e.g., Demir-Lira et al., 2018; Dick et al., 2014; Holler et al., 2015; Kircher et al., 2009; Straube et al., 2011; Willems et al., 2007). Other claims have focused on gestures that are independently interpretable (i.e., emblems; e.g., Andric et al., 2013; Enrici, Adenzato, Cappa, Bara, & Tettamanti, 2011; Xu, Gannon, Emmorey. Smith, & Braun, 2009; Redcay et al., 2016; Papeo, Agostini, & Lingnau, in press). Our results do not directly speak to these claims as our study contained a relatively small proportion of emblematic gestures (~16%) and a small proportion of gestures that clarify the linguistic message (~5%). We did not examine responses to these different kinds of gestures separately.

The results of our Behavioral Study 4, where we had an independent group of participants make guesses about the content of the accompanying speech signal for the silent videos of gesture movements, showed that the gesture movements contain a substantial amount of information (as evidenced by relatively high lexical overlap in the participants' productions). So, what our study clearly demonstrates is that brain regions that support language comprehension do not respond to diverse naturalistically produced co-speech gestures in the absence of speech, despite the fact that these are rich communicative signals that often carry

information that reinforces or supplements the linguistic signal. Instead, these language-processing regions appear to be specialized for processing linguistic input (e.g., Fedorenko et al., 2011; Fedorenko & Varley, 2016). Although it remains possible that some language areas are sensitive to language-gesture integration for specific types of gestures (Figure 1C/D), or to emblematic gestures even in the absence of speech (Figure 1A), the limitations outlined above, which characterize many prior studies, should be taken into account when interpreting their results.

Our results stand in apparent contrast to a recent study that used an individual fROI approach, similar to the one used here, and reported overlapping responses to language and gesture within left STS (Redcay et al., 2016). However, the effect sizes for the responses to linguistic stimuli vs. gestures were not reported in that study. As a result, we believe the results in Redcay et al. may be similar to ours in spite of different interpretations. Recall that our analysis of overlapping voxels revealed overlap in a substantial proportion of participants within posterior temporal cortex (a broadly similar area to the one reported by Redcay et al.). However, the magnitude of the response to the *SilentGesture* condition was close to the fixation baseline, and the fine-grained pattern of activation was distinct from that elicited by language comprehension (and not even stable across runs). These results strongly argue against the idea that the same mechanisms are used to process linguistic information and non-linguistic communicative signals like gestures.

Although the language regions did not respond differentially to gestures vs. the control (grooming) movements, participants could clearly distinguish the two types of movements behaviorally (Behavioral Study 1), including judging gestures to be significantly more communicative than grooming movements (Behavioral Study 3). Further, the two conditions elicited differential response magnitudes in areas outside of the language network. In particular, gestures elicited a stronger response than grooming movements bilaterally in the posterior temporal cortex, as well as some frontal areas and the right cerebellum. The proximity of the gesture-responsive area in the left temporal cortex to the language-responsive areas, and in the right temporal cortex to the right-hemisphere homologs of the language areas, is intriguing. Given the general topographic similarity between the language brain areas and areas that support social cognition (e.g., Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Paunov, Blank, & Fedorenko, 2019), it seems possible that lateral frontal and temporal areas had housed mechanisms for processing any communicative signals in our ancestors. Indeed, in a recent fMRI study with macaques, Shepherd & Freiwald (2018) reported a set of frontal and temporal areas that respond to the interpretation of socially-relevant signals, and whose topography grossly resembles the language network in humans. However, in human brains, lateral frontal and temporal cortex expanded substantially and apparently fractionated into a myriad of functionally distinct areas, with some areas being selective for linguistic processing (e.g., Fedorenko et al., 2011), and others being selective for processing different non-linguistic communicative signals, like facial expressions (e.g., Pitcher et al., 2011), body postures (e.g., Downing, Peelen, Wiggett, & Tew, 2006), voices (e.g., Belin, Fecteau, & Bedard, 2004; cf. Norman-Haignere, Kanwisher, & McDermott, 2015), prosody (e.g., Fedorenko, Hsieh, & Balewski, 2015), etc.

Gestures also elicited quite strong responses in parts of the domain-general fronto-parietal multiple demand (MD) network. However, although the MD brain regions may support gesture processing, our data suggest that they support action processing more generally, rather than gesture specifically. In particular, MD responses were not selective for gestures relative to grooming movements, in line with prior work implicating parts of the MD network in action observation (e.g., Biagi et al., 2015; Caspers et al., 2010; Culham & Valyear, 2006; Gallivan & Culham, 2015). It is also important to keep in mind that the MD network is sensitive to effort across diverse cognitive tasks (e.g., Duncan & Owen, 2001; Fedorenko et al., 2013; Hugdahl et al., 2015). Responses in the MD regions (defined in the current study by a demanding spatial working memory task) to gesture may therefore reflect general executive demands rather than action-specific computations. This is especially relevant in interpreting results from gesture studies that manipulate difficulty. For example, several studies have reported activation within left IPS when participants processed gestures that were either unrelated to (Green et al., 2009) or incongruent with (Willems et al., 2007) the accompanying speech. The most parsimonious explanation of these effects is in terms of general (attention/cognitive control) processes. To argue for a gesture- or action-specific interpretation, it would be necessary to show that grooming movements, or non-action related demanding tasks, do not elicit similar responses in the same regions.

Gesture and speech form an integrated system at the behavioral level (e.g., by taking into account the information conveyed in both speech *and* gesture during novel task performance, we can more reliably infer the learner's degree of competence; Church & Goldin-Meadow, 1986; Perry, Church & Goldin-Meadow, 1988; Goldin-Meadow, Alibali, & Church, 1993; for review, see Goldin-Meadow, 2003). Our findings suggest, however, that this integrated behavioral system is *not* implemented within the same neural mechanisms. Rather, the linguistic signal and co-speech gestures appear to be processed by distinct brain regions, even though some of these regions appear to lie in close proximity to each other within left posterior temporal cortex.

To conclude, we demonstrated that brain regions that support language comprehension are functionally specialized for processing linguistic signals. Gestures – non-verbal communicative signals that accompany speech – appear to be processed in brain areas outside of the language network, much like emotional prosody (e.g., Ross, 1981), facial expressions (e.g., Pitcher et al., 2011) and eye gaze (e.g., Itier & Batty, 2009). Our results therefore show that language processing and gesture processing rely on distinct cognitive and neural resources. This functional separability is not, however, inconsistent with strong integration between linguistic and non-linguistic communicative signals in online processing. For example, we robustly integrate information about people's faces and bodies although their processing is supported by distinct regions within the ventral visual stream (e.g., Schwarzlose, Baker, & Kanwisher, 2005). How exactly the integration of the linguistic signal with non-verbal communicative signals is implemented neurally remains to be discovered. Inter-region synchronization of neural activity may provide one possible mechanism (e.g., Paunov et al., 2019).

Acknowledgements:

We thank Chris Goodnow, Julie Hwang, Ryan Kang, Daniel Song, Victoria Su, Stephanie Tam, Ming-yee Tsang, and Stephanie Velazquez for their help with the creation of the gesture materials; Leyla Isik and Sarah Schwettman for their advice regarding computing motion information; and Hope Kean for help with the figures. We would like to acknowledge the Athinoula A. Martinos Imaging Center at McGovern Institute for Brain Research at MIT, and the support team (Steve Shannon, Atsushi Takahashi, and Sheeba Arnold). This work was supported by a K99R00 award HD057522 from NICHD and an R01 award DC016607 from NIDCD to EF, a grant from the Simons Foundation through the Simons Center for the Social Brain at MIT to EF, grants ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX), and funds from the Center for Gesture, Sign and Language at the University of Chicago.

References

Abner N, Cooperrider K, & Goldin-Meadow S (2015). Gesture for linguists: A handy primer. Language and Linguistics Compass. 9, 437–449. [PubMed: 26807141]

Adolphs R (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. Behavioral and cognitive neuroscience reviews, 1(1), 21–62. [PubMed: 17715585]

Amit E, Hoeflin C, Hamzah N, & Fedorenko E (2017). An asymmetrical relationship between verbal and visual thinking: Converging evidence from behavior and fMRI. NeuroImage, 152, 619–627. [PubMed: 28323162]

Andric M, Solodkin A, Buccino G, Goldin-Meadow S, Rizzolatti G, & Small SL (2013). Brain function overlaps when people observe emblems, speech, and grasping. Neuropsychologia. 51, 1619–1629. [PubMed: 23583968]

Bautista A, & Wilson SM (2016). Neural responses to grammatically and lexically degraded speech. Language, cognition and neuroscience, 31(4), 567–574.

Belin P, Fecteau S, & Bedard C (2004). Thinking the voice: neural correlates of voice perception. Trends in cognitive sciences, 8(3), 129–135. [PubMed: 15301753]

Benjamini Y, & Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics. 1, 1165–1188.

Biagi L, Cioni G, Fogassi L, Guzzetta A, Sgandurra G, & Tosetti M (2016). Action observation network in childhood: A comparative fMRI study with adults. Developmental Science. 19, 1075 [PubMed: 26537750]

Binder JR, Frost JA, Hammeke TA, Cox RW, Rao SM, & Prieto T (1997). Human brain language areas identified by functional magnetic resonance imaging. Journal of Neuroscience, 17(1), 353–362. [PubMed: 8987760]

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

Buck R (1984). The physiological bases of nonverbal communication In Sociophysiology (pp. 139–161). Springer, New York, NY.

Buck R, & VanLear CA (2002). Verbal and nonverbal communication: Distinguishing symbolic, spontaneous, and pseudo-spontaneous nonverbal behavior. Journal of communication, 52(3), 522–541.

Campbell N (2007). On the use of nonverbal speech sounds in human communication In Verbal and nonverbal communication behaviours (pp. 117–128). Springer, Berlin, Heidelberg.

Caspers S, Zilles K, Laird AR, & Eickhoff SB (2010). ALE meta-analysis of action observation and imitation in the human brain. NeuroImage, 50, 1148–1167. [PubMed: 20056149]

Chai LR, Mattar MG, Blank IA, Fedorenko E and Bassett DS (2016). Functional network dynamics of the language system. Cerebral Cortex, 26, 4148–4159. [PubMed: 27550868]

Chang L, & Tsao DY (2017). The code for facial identity in the primate brain. Cell, 169(6), 1013–1028. [PubMed: 28575666]

Chen G, Taylor PA, & Cox RW (2017). Is the statistic value all we should care about in neuroimaging?. Neuroimage, 147, 952–959. [PubMed: 27729277]

Church RB, & Goldin-Meadow S (1986). The mismatch between gesture and speech as an index of transitional knowledge. Cognition, 23, 43–71. [PubMed: 3742990]

Cover TM, & Thomas JA (1991). Entropy, relative entropy and mutual information. Elements of information theory, 2, 1–55.

Culham JC, & Valyear KF (2006). Human parietal cortex in action. Current opinion in neurobiology, 16, 205–212. [PubMed: 16563735]

Cutler A, Dahan D, & Van Donselaar W (1997). Prosody in the comprehension of spoken language: A literature review. Language and speech, 40(2), 141–201. [PubMed: 9509577]

de Beer C, Carragher M, van Nispen K, Hogrefe K, de Ruiter JP, & Rose ML (2017). How much information do people with aphasia convey via gesture?. American Journal of Speech-Language Pathology, 26(2), 483–497. [PubMed: 28492911]

Deen B, Koldewyn K, Kanwisher N, & Saxe R (2015). Functional organization of social perception and cognition in the superior temporal sulcus. Cerebral Cortex, 25, 4596–4609. [PubMed: 26048954]

Demir-Lira OE, Asaridou S, Beharelle AR, Holt A, Goldin-Meadow S, Small S (2018). Functional neuroanatomy of gesture-speech integration in children varies with individual differences in gesture processing. Developmental Science.

Dick AS, Goldin-Meadow S, Hasson U, Skipper JI, & Small SL (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. Human brain mapping, 30, 3509–3526. [PubMed: 19384890]

Dick AS, Goldin-Meadow S, Solodkin A, & Small SL (2012). Gesture in the developing brain. Developmental Science, 15, 165–180. [PubMed: 22356173]

Dick AS, Mok EH, Beharelle AR, Goldin-Meadow S, & Small SL (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. Human Brain Mapping, 35, 900–917. [PubMed: 23238964]

Dodell-Feder D, Koster-Hale J, Bedny M, & Saxe R (2011). fMRI item analysis in a theory of mind task. Neuroimage, 55(2), 705–712. [PubMed: 21182967]

Downing PE, Peelen MV, Wiggett AJ, & Tew BD (2006). The role of the extrastriate body area in action perception. Social Neuroscience, 1(1), 52–62. [PubMed: 18633775]

Duncan J (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends in cognitive sciences, 14, 172–179. [PubMed: 20171926]

Duncan J (2013). The structure of cognition: attentional episodes in mind and brain. Neuron, 80, 35–50. [PubMed: 24094101]

Duncan J, & Owen AM (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends in Neurosciences, 23, 475–483. [PubMed: 11006464]

Eggenberger N, Preisig BC, Schumacher R, Hopfner S, Vanbellingen T, Nyffeler T, … & Muri RM. (2016). Comprehension of co-speech gestures in aphasic patients: an eye movement study. PloS one, 11(1), e0146583. [PubMed: 26735917]

Ekman P, & Friesen WV (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica, 1, 49–98.

Enrici I, Adenzato M, Cappa S, Bara BG, & Tettamanti M (2011). Intention processing in communication: a common brain network for language and gestures. Journal of Cognitive Neuroscience, 23, 2415–2431. [PubMed: 20954937]

Fedorenko E, Behr MK, & Kanwisher N (2011). Functional specificity for high-level linguistic processing in the human brain. PNAS, 108, 16428–16433. [PubMed: 21885736]

Fedorenko E, Duncan J, & Kanwisher N (2013). Broad domain generality in focal regions of frontal and parietal cortex. PNAS, 110, 16616–21. [PubMed: 24062451]

Fedorenko E, Duncan J & Kanwisher N (2012). Language-selective and domain-general regions lie side by side within Broca's area. Current Biology, 22, 2059–2062. [PubMed: 23063434]

Fedorenko E, Hsieh PJ, & Balewski Z (2015). A possible functional localiser for identifying brain regions sensitive to sentence-level prosody. Language, cognition and neuroscience, 30, 120–148.

Fedorenko E, Hsieh P-J, Nieto-Castañón A, Whitfield-Gabrieli S, & Kanwisher N (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. Journal of Neurophysiology, 104, 1177–1194. [PubMed: 20410363]

Fedorenko E, Nieto-Castaon A, & Kanwisher N (2012a). Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. Neuropsychologia, 50(4), 499–513. [PubMed: 21945850]

Fedorenko E, & Varley R (2016). Language and thought are not the same thing: evidence from neuroimaging and neurological patients. Annals of the New York Academy of Sciences, 1369, 132–153. [PubMed: 27096882]

Fedorenko E et al. (2018). Word meanings and sentence structure recruit the same set of fronto-temporal regions during comprehension. bioRxiv 477851

Fisher RA (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron, 1, 3–32.

Fleiss JL, & Cohen J (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 33(3), 613–619.

Friederici AD, Meyer M, & von Cramon DY (2000). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. Brain and Language, 74, 289–300. [PubMed: 10950920]

Friston KJ, Rotshtein P, Geng JJ, Sterzer P, & Henson RN (2006). A critique of functional localisers. NeuroImage, 30, 1077–1087. [PubMed: 16635579]

Frost MA, Goebel R, (2011). Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage, 59, 1369–1381. [PubMed: 21875671]

Gallivan JP, & Culham JC (2015). Neural coding within human brain areas involved in actions. Current opinion in neurobiology, 33, 141–149. [PubMed: 25876179]

Glezer LS, & Riesenhuber M (2013). Individual variability in location impacts orthographic selectivity in the “visual word form area”. Journal of Neuroscience, 33, 11221–11226. [PubMed: 23825425]

Gobl C, & Chasaide AN (2003). The role of voice quality in communicating emotion, mood and attitude. Speech communication, 40(1-2), 189–212.

Goldberg AE (1995). Constructions: A construction grammar approach to argument structure. University of Chicago Press.

Goldin-Meadow S (2003). Hearing gesture: How our hands help us think. Harvard University Press.

Goldin-Meadow S, Alibali MW, & Church RB (1993). Transitions in concept acquisition: Using the hand to read the mind. Psychological Review, 100, 279–297. [PubMed: 8483984]

Green A, Straube B, Weis S, Jansen A, Willmes K, Konrad K, & Kircher T (2009). Neural integration of iconic and unrelated coverbal gestures: A functional MRI study. Human Brain Mapping, 30, 3309–3324. [PubMed: 19350562]

Hage SR, & Nieder A (2016). Dual neural network model for the evolution of speech and language. Trends in neurosciences, 39(12), 813–829. [PubMed: 27884462]

Harper RG, Wiens AN, & Matarazzo JD (1978). Nonverbal communication: The state of the art. John Wiley & Sons.

He Y, Gebhardt H, Steines M, Sammer G, Kircher T, Nagels A, & Straube B (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. Neuropsychologia, 72, 27–42. [PubMed: 25900470]

Holle H, Gunter TC, Rueschemeyer SA, Hennenlotter A, & Iacoboni M (2008). Neural correlates of the processing of co-speech gestures. NeuroImage, 39, 2010–2024. [PubMed: 18093845]

Holler J, Kokal I, Toni I, Hagoort P, Kelly SD, & Özyürek A (2015). Eye’m talking to you: Speakers’ gaze direction modulates co-speech gesture processing in the right MTG. Social Cognitive and Affective Neuroscience, 10, 255–261. [PubMed: 24652857]

Hubbard AL, Wilson SM, Callan DE, & Dapretto M (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. Human Brain Mapping, 30, 1028–1037. [PubMed: 18412134]

Hugdahl K, Raichle ME, Mitra A, & Specht K (2015). On the existence of a generalized non-specific task-dependent network. Frontiers in human neuroscience, 9, 430. [PubMed: 26300757]

Humphries C, Binder JR, Medler DA, & Liebenthal E (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. Journal of cognitive neuroscience, 18, 665–679. [PubMed: 16768368]

Itier RJ, & Batty M (2009). Neural bases of eye and gaze processing: the core of social cognition. Neuroscience & Biobehavioral Reviews, 33(6), 843–863. [PubMed: 19428496]

Josse G, Joseph S, Bertasi E, Giraud AL (2012). The brain's dorsal route for speech represents word meaning: evidence from gesture. PLoS ONE, 7, e46108. [PubMed: 23049951]

Julian JB, Fedorenko E, Webster J, & Kanwisher N (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. NeuroImage, 60, 2357–2364. [PubMed: 22398396]

Kanwisher N, McDermott J, & Chun MM (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience, 17(11), 4302–4311. [PubMed: 9151747]

Kendon A (1967). Some functions of gaze-direction in social interaction. Acta psychologica, 26, 22–63. [PubMed: 6043092]

Kendon A (1980). Gesticulation and speech: Two aspects of the process of utterance, in The relationship of verbal and nonverbal communication, 25, 207–227.

Kircher T, Straube B, Leube D, Weis S, Sachs O, Willmes K, … Green A. (2009). Neural interaction of speech and gesture: Differential activations of metaphoric co-verbal gestures. Neuropsychologia, 47, 169–179. [PubMed: 18771673]

Kita S, Gijn IV, & van der Hulst H (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. Gesture and sign language in human-computer interaction, 23–35.

Mahowald K, & Fedorenko E (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. NeuroImage, 139, 74–93. [PubMed: 27261158]

Marstaller L, & Burianová H (2014). The multisensory perception of co-speech gestures - A review and meta-analysis of neuroimaging studies. Journal of Neurolinguistics, 30, 69–77.

McNeill D (1992). Hand and mind: What gestures reveal about thought. University of Chicago Press.

McNeill D (2000). Language and Gesture. Language, Culture & Cognition.

Mineroff Z, Blank IA, Mahowald K, & Fedorenko E (2018). A robust dissociation among the language, multiple demand, and default mode networks: Evidence from inter-region correlations in effect size. Neuropsychologia, 119, 501–511. [PubMed: 30243926]

Müller C, Cienki A, Fricke E, Ladewig S, McNeill D, & Tessendorf S (Eds.). (2013). Body-Language-Communication (Vol. 1). Walter de Gruyter.

Nagels A, Chatterjee A, Kircher T, Straube B (2013). The role of semantic abstractness and perceptual category in processing speech accompanied by gestures. Frontiers Behavioral Neuroscience, 7, 181.

Nieto-Castañon A, & Fedorenko E (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. NeuroImage, 63, 1646–1669. [PubMed: 22784644]

Norman-Haignere S, Kanwisher NG, & McDermott JH (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron, 88(6), 1281–1296. [PubMed: 26687225]

Novack MA, & Goldin-Meadow S (2017). Gesture as representational action: A paper about function. Psychonomic bulletin & review, 24(3), 652–665. [PubMed: 27604493]

Oldfield RC 1971 The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9, 97–113. [PubMed: 5146491]

Papeo L, Agostini B, Lingnau A (in press). The large-scale organization of gestures and words in the middle temporal gyrus. Journal of Neuroscience.

Paunov AM, Blank IA, & Fedorenko E (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. Journal of neurophysiology, 121(4), 1244–1265. [PubMed: 30601693]

Perry M, Church RB, & Goldin-Meadow S (1988). Transitional knowledge in the acquisition of concepts. Cognitive Development, 3, 359–400.

Pitcher D, Dilks DD, Saxe RR, Triantafyllou C, & Kanwisher N (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. NeuroImage, 56(4), 2356–2363. [PubMed: 21473921]

Poldrack RA (2006). Can cognitive processes be inferred from neuroimaging data?. Trends in cognitive sciences, 10(2), 59–63. [PubMed: 16406760]

Pritchett BL, Hoeflin C, Koldewyn K, Dechter E, & Fedorenko E (2018). High-level language processing regions are not engaged in action observation or imitation. Journal of neurophysiology, 120(5), 2555–2570. [PubMed: 30156457]

Rauschecker JP (2018). Where did language come from? Precursor mechanisms in nonhuman primates. Current opinion in behavioral sciences, 21, 195–204. [PubMed: 30778394]

Redcay E, Velnoskey KR, & Rowe ML (2016). Perceived communicative intent in gesture and language modulates the superior temporal sulcus. Human brain mapping. 37, 3444–3461 [PubMed: 27238550]

Rogalsky C, & Hickok G (2011). The role of broca's area in sentence comprehension. Journal of Cognitive Neuroscience, 23, 1664–1680. [PubMed: 20617890]

Ross ED (1981). The aprosodias: Functional-anatomic organization of the affective components of language in the right hemisphere. Archives of Neurology, 38(9), 561–569. [PubMed: 7271534]

Sato S, & Kawahara JI (2015). Attentional capture by completely task-irrelevant *faces*. Psychological research, 79, 523–533. [PubMed: 25030814]

Saxe R, Brett M, & Kanwisher N (2006). Divide and conquer: a defense of functional localizers. NeuroImage, 30, 1088–1096. [PubMed: 16635578]

Schwarzlose RF, Baker CI, & Kanwisher N (2005). Separate face and body selectivity on the fusiform gyrus. Journal of Neuroscience, 25, 11055–11059. [PubMed: 16306418]

Scott TL, Gallée J, & Fedorenko E (2016). A new fun and robust version of an fMRI localizer for the frontotemporal language system. Cognitive Neuroscience, 8(3), 167–176. [PubMed: 27386919]

Shepherd SV, & Freiwald WA (2018). Functional networks for social communication in the macaque monkey. Neuron, 99(2), 413–420. [PubMed: 30017395]

Silver NC, & Dunlap WP (1987). Averaging correlation coefficients: should Fisher's z transformation be used?. Journal of applied psychology, 72(1), 146.

Skipper JI, Goldin-Meadow S, Nusbaum HC, & Small SL (2007). Speech-associated gestures, Broca's area, and the human mirror system. Brain and Language, 101, 260–277. [PubMed: 17533001]

Straube B, Green A, Bromberger B, & Kircher T (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. Human Brain Mapping, 32, 520–533. [PubMed: 21391245]

Straube B, Green A, Weis S, Kircher T (2012). A supramodal neural network for speech and gesture semantics: an fMRI study. PLoS ONE, 7, e51207. [PubMed: 23226488]

Sullivan GM, & Feinn R (2012). Using effect size—or why the P value is not enough. Journal of graduate medical education, 4(3), 279–282. [PubMed: 23997866]

Tahmasebi AM, Davis MH, Wild CJ, Rodd JM, Hakyemez H, Abolmaesumi P, & Johnsrude IS (2012). Is the link between anatomical structure and function equally strong at all cognitive levels of processing?. Cerebral Cortex, 22, 1593–1603. [PubMed: 21893681]

Thesen S, Heid O, Mueller E, & Schad LR (2000). Prospective Acquisition Correction for head motion with image-based tracking for real-time fMRI. Magnetic Resonance in Medicine, 44, 457–465. [PubMed: 10975899]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, … & Joliot M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage, 15, 273–289. [PubMed: 11771995]

Von Humboldt W (1836/1999). Humboldt: 'On Language': On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species. Cambridge University Press.

Weintraub S, Mesulam MM, & Kramer L (1981). Disturbances in prosody: A right-hemisphere contribution to language. Archives of Neurology, 38(12), 742–744. [PubMed: 7316838]

Weisberg J, Hubbard AL, & Emmorey K (2016). Multimodal integration of spontaneously produced representational co-speech gestures: An fMRI study. Language, Cognition and Neuroscience, 32, 158–174.

Willems RM, Ozyürek A, & Hagoort P (2007). When language meets action: The neural integration of gesture and speech. Cerebral Cortex, 17, 2322–2333. [PubMed: 17159232]

Willems RM, & Hagoort P (2007). Neural evidence for the interplay between language, gesture, and action: A review. Brain and Language, 101, 278–289. [PubMed: 17416411]

Willems RM, Ozyürek A, & Hagoort P (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. NeuroImage, 47, 1992–2004. [PubMed: 19497376]

Xu J, Gannon PJ, Emmorey K, Smith JF, & Braun AR (2009). Symbolic gestures and spoken language are processed by a common neural system. PNAS, 106, 20664–20669. [PubMed: 19923436]

Yang J, Andric M, & Mathew MM (2015). The neural basis of hand gesture comprehension: A meta-analysis of functional magnetic resonance imaging studies. Neuroscience and Biobehavioral Reviews, 57, 88–104. [PubMed: 26271719]

Highlights

- Some have argued that processing gestures engages the brain regions that support language comprehension.
- We used fMRI to examine responses to gesture processing in language regions defined functionally in individual participants, including directly comparing effect sizes, and covering a broad range of spontaneously generated co-speech gestures.
- Whenever speech was present, language regions responded robustly (and to a similar degree regardless of whether the video contained gestures or grooming movements).
- In contrast, and critically, responses in the language regions were low when silent videos were processed (again, regardless of whether they contained gestures or grooming movements).
- Brain regions outside of the language network, including some in close proximity to its regions, differentiated between gestures and grooming movements.
- In summary, contra prior claims, language-processing regions do not respond to co-speech gestures in the absence of speech, suggesting that these regions are selectively driven by linguistic input (e.g., Fedorenko et al., 2011).

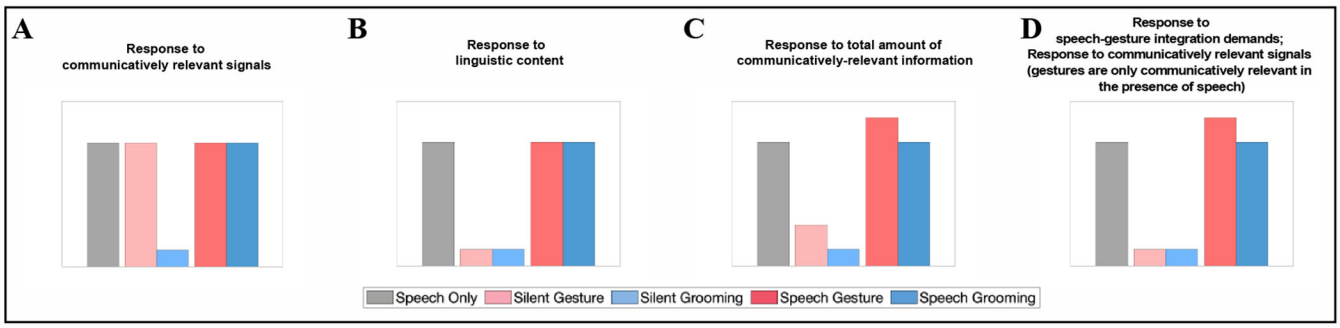

Figure 1:

Predicted patterns of responses of the language fROIs to the conditions of the gesture experiment. **A.** Response to communicatively relevant signals. **B.** Response to linguistic content. **C.** Response to total amount of communicatively-relevant information. **D.** Response to speech-gesture integration demand + communicatively-relevant information, but gestures are only communicatively relevant in the presence of speech.

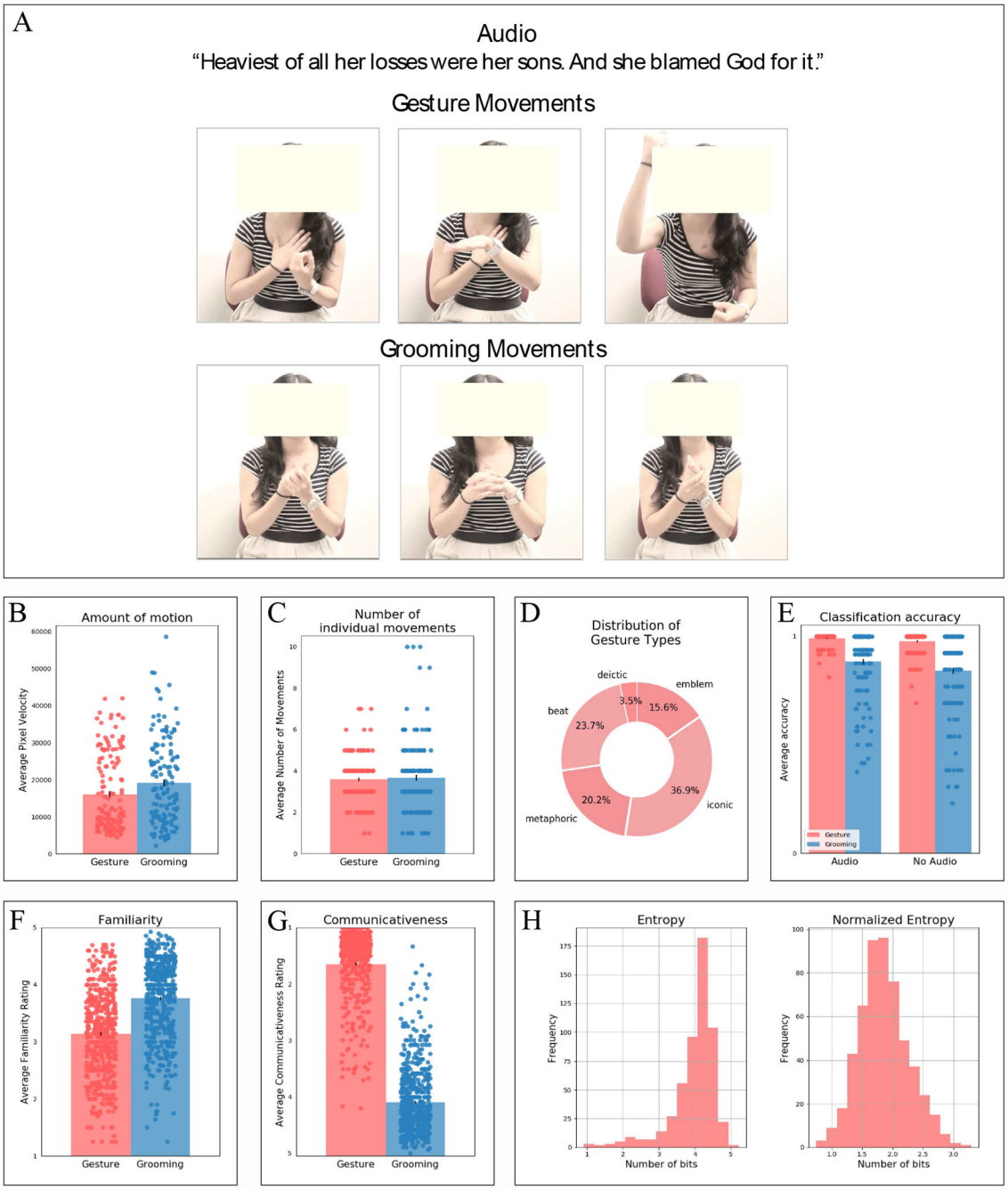

Figure 2:

A. Stills from the stimuli in the critical gesture experiment. **B.** Amount of motion in the gesture and grooming videos. **C.** Number of individual movements in the gesture and grooming videos. **D.** Distribution of gesture types in the gesture videos. **E.** Classification accuracy for the gesture and grooming videos presented with or without audio. **F.** Familiarity ratings of the gesture and grooming movements. **G.** Communicativeness ratings for the gesture and grooming movements. **H**. Distribution of entropy values and normalized entropy values for the gesture movements. NB: In Figures 2B, C, and E, the dots correspond to

individual gesture or grooming videos; in Figures 2F and G, the dots correspond to individual gesture or grooming movements.

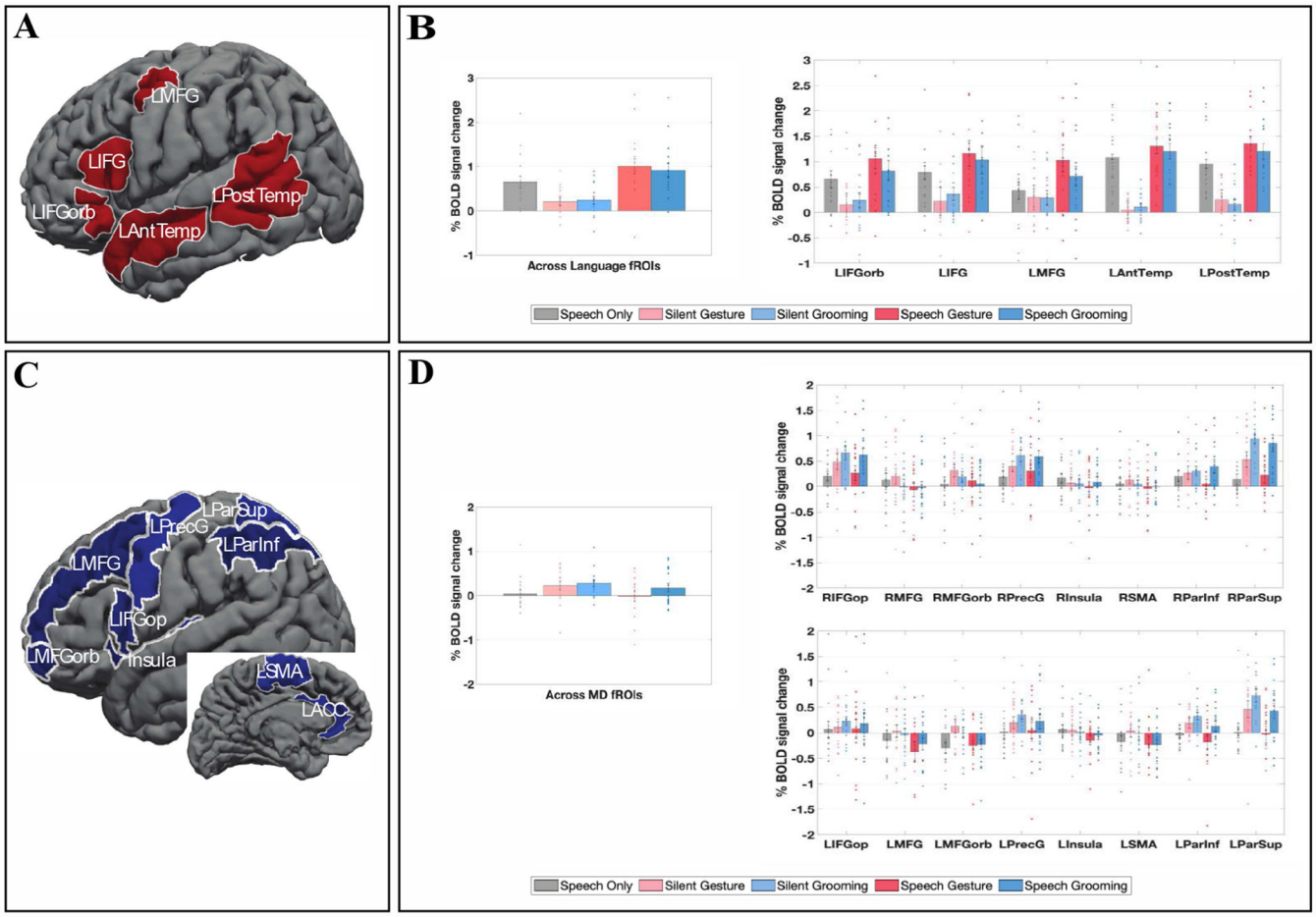

Figure 3.

A. The parcels used to define the Language fROIs. **B.** Responses of the language fROIs to the conditions of the gesture experiment. Error bars correspond to standard errors of the mean. Dots correspond to responses of individual participants. **C.** The parcels used to define the MD fROIs. **D.** Responses of the MD fROIs to the conditions of the gesture experiment. Error bars correspond to standard errors of the mean. Dots correspond to responses of individual participants.

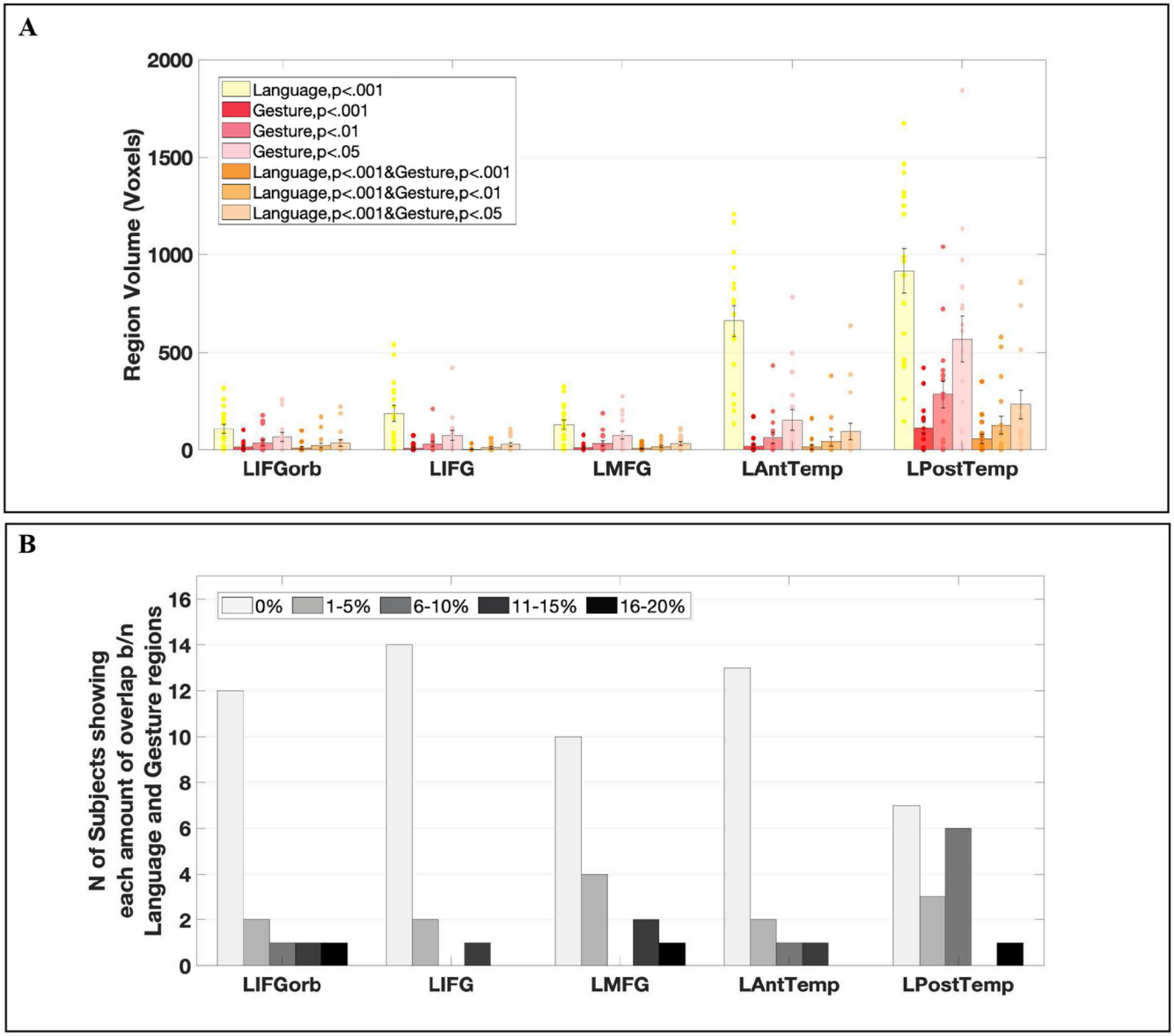

Figure 4.

A. The volumes of the language fROIs (defined by the Sentence > Nonwords contrast) and gesture fROIs (defined by the Gesture > Grooming contrast) within the six language parcels (Figure 2, left). Dots correspond to responses of individual participants. **B.** The number of subjects showing different amounts of overlap (as quantified by the Jaccard index, i.e., the number of voxels that overlap between language and gesture regions divided by the total number of voxels for both contrasts, multiplied by 100)) between language- and gesture-responsive voxels, significant at the whole-brain threshold level of $p < 0.001$, within the language parcels (mean overlap across subjects across parcels = 3%, SD =4%)

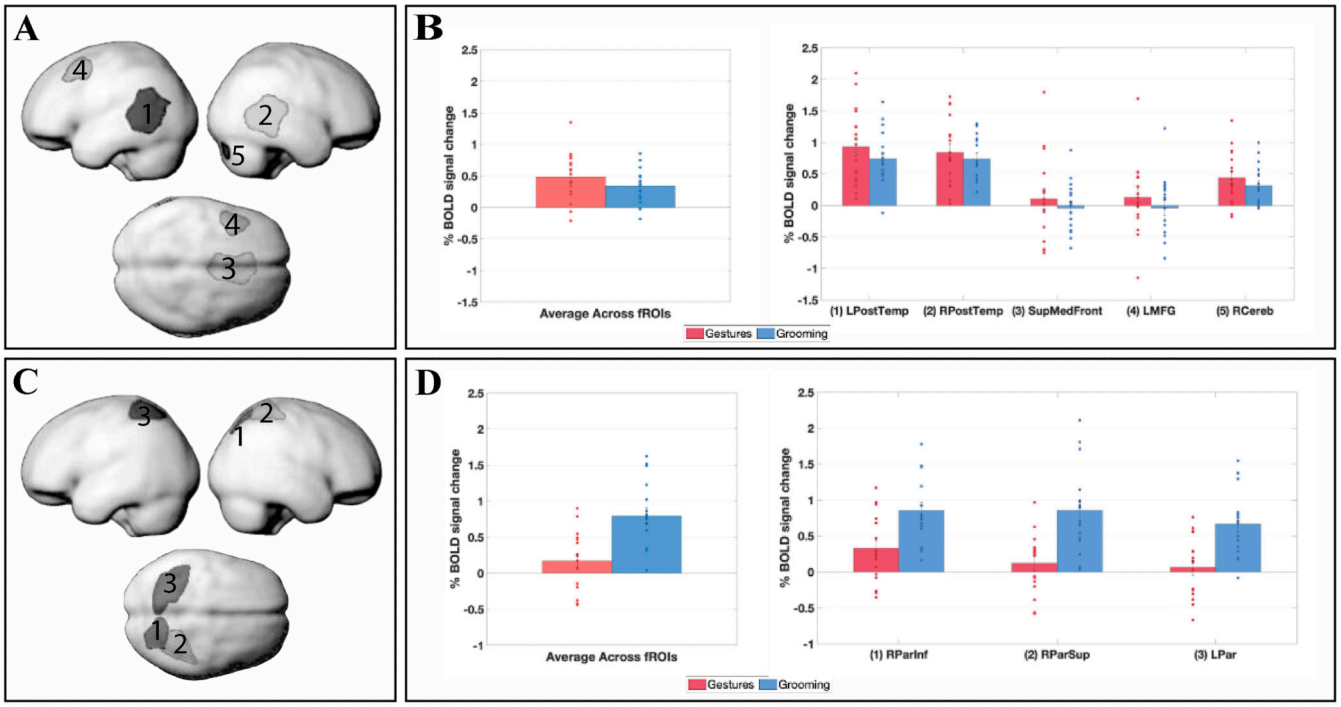

Figure 5.

A. Brain regions sensitive to the Gesture > Grooming contrast as identified by the whole-brain GSS analysis. **B**. Responses of the Gesture > Grooming fROIs to gestures and grooming movements. Error bars correspond to standard errors of the mean. Dots correspond to responses of individual participants. **C**. Brain regions sensitive to the Grooming > Gesture contrast as identified by a whole-brain GSS analysis. **D.** Responses of the Grooming > Gesture fROIs to gestures and grooming movements. Error bars correspond to standard errors of the mean. Dots correspond to responses of individual participants.

Table 1.

Summary of paired-samples t-tests. Uncorrected p values (df = 1,16) are reported. Effects that remain significant after an FDR correction for the number of regions (language fROIs = 5; MD fROIs = 18) are shown in bold.

	fROIs	SilentGesture vs. SilentGrooming	SilentGesture vs. SpeechOnly	SpeechGesture vs. SpeechOnly	SpeechGesture vs. SpeechGrooming
LANGUAGE	LIFGorb	*t*=−.72, *p*=.48	***t*=−4.10, *p*=.001**	***t*=5.03, *p*<.001**	*t*=1.51, *p*=.15
	LIFG	*t*=−1.35, *p*=.20	***t*=−3.17, *p*=.006**	***t*=4.46, *p*<.001**	*t*=.82, *p*=.43
	LMFG	*t*=.07, *p*=.95	***t*=−2.67, *p*=.02**	***t*=3.72, *p*=.002**	*t*=1.60, *p*=.13
	LAntTemp	*t*=−1.05, *p*=.31	***t*=−7.77, *p*<.001**	***t*=8.04, *p*<.001**	*t*=1.14, *p*=.27
	LPostTemp	*t*=1.71, *p*=.11	***t*=−5.27, *p*<.001**	***t*=6.40, *p*<.01**	*t*=1.21, *p*=.24
MD	LIFGorb	*t*=−1.19, *p*=.25	*t*=.35, *p*=.73	*t*=.04, *p*=.97	*t*=−.67, *p*=.52
	RIFGorb	*t*=−1.33, *p*=.20	*t*=1.58, *p*=.13	*t*=.03, *p*=.98	*t*=−1.75, *p*=.10
	LMFG	*t*=.56, *p*=.59	*t*=2.04, *p*=.06	*t*=−2.44, *p*=.03	*t*=−.85, *p*=.41
	RMFG	*t*=1.13, *p*=.28	*t*=.63, *p*=.54	*t*=−1.45, *p*=.17	*t*=−.32, *p*=.76
	LMFGorb	*t*=1.44, *p*=.17	*t*=2.65, *p*=.02	*t*=−2.30, *p*=.04	*t*=−.06, *p*=.96
	RMFGorb	*t*=.69, *p*=.50	*t*=1.84, *p*=.09	*t*=−.80, *p*=.44	*t*=.43, *p*=.68
	LPrecG	*t*=−1.45, *p*=.17	*t*=1.37, *p*=.19	*t*=−.41, *p*=.69	*t*=−.80, *p*=.43
	RPrecG	*t*=−2.39, *p*=.03	*t*=1.30, *p*=.21	*t*=.82, *p*=.42	*t*=−.1.74, *p*=. 10
	L_Insula	*t*=.30, *p*=.77	*t*=.83, *p*=.42	*t*=−1.98, *p*=.07	*t*=−.73, *p*=.48
	R_Insula	*t*=.03, *p*=.98	*t*=−.07, *p*=.95	*t*=−.76, *p*=.46	*t*=−.80, *p*=.45
	LSMA	*t*=.41, *p*=.69	*t*=1.48, *p*=.16	*t*=−1.43, *p*=.17	*t*=.10, *p*=.92
	RSMA	*t*=.78, *p*=.45	*t*=.78, *p*=.45	*t*=−1.23, *p*=.24	*t*=−.25, *p*=.80
	LParInf	*t*=−1.61, *p*=.13	*t*=2.81, *p*=.01	*t*=−1.76, *p*=.10	*t*=−2.10, *p*=.05
	RParInf	*t*=−.48, *p*=.64	*t*=.86, *p*=.41	*t*=−.46, *p*=.65	*t*=−2.45, *p*=.02
	LParSup	*t*=−2.66, *p*=.02	*t*=3.20, *p*=.006	*t*=−1.67, *p*=.11	*t*=−2.61, *p*=.02
	RParSup	*t*=−2.97, *p*=.009	*t*=2.93, *p*=.01	*t*=−.02, *p*=.99	*t*=−3.28, *p*=.005
	LACC	*t*=1.77, *p*=.09	*t*=2.05, *p*=.06	*t*=−2.92, *p*=.01	*t*=−44, *p*=.66
	RACC	*t*=2.17, *p*=.05	*t*=1.48, *p*=.16	*t*=−2.70, *p*=.02	*t*=.12, *p*=.90

Table 2.

Brain regions sensitive to the Gesture > Grooming contrast as identified by a whole-brain GSS analysis. Effects that remain significant after an FDR correction for the number of regions are shown in bold.

ROI #	ROI size	Approximate anatomical location	Proportion of subjects in whom the fROI is present	*t*-value for the Gesture > Grooming contrast
1	226	L Posterior Temporal	.88	**3.05**
2	170	R Posterior Temporal	.71	**5.52**
3	169	Medial Frontal Cortex	.71	**3.25**
4	59	L Superior Frontal Gyrus	.71	**2.30**
5	32	R Cerebellum	.71	**3.52**

Table 3.

Brain regions sensitive to the Grooming > Gesture contrast as identified by a whole-brain GSS analysis. Effects that remain significant after an FDR correction for the number of regions are shown in bold.

ROI #	ROI size	Approximate anatomical location	Proportion of subjects in whom the fROI is present	*t*-value for the Gesture > Grooming contrast
1	129	R Superior Parietal	.71	**3.90**
2	146	R Inferior Parietal	.71	**9.38**
3	240	L Superior Parietal	.71	**7.50**