Representation of action concepts in left posterior temporal cortex that generalize across vision and language

Moritz F. Wurm[1], Alfonso Caramazza[1,2]

[1] Center for Mind/Brain Sciences (CIMeC), University of Trento, Italy
[2] Department of Psychology, Harvard University, USA

Corresponding author: Moritz Wurm, Center for Mind/Brain Sciences (CIMeC), University of Trento, Corso Bettini 31, 38068 Rovereto, Italy. Phone: +39 0461-28 8729, Email: moritz.wurm@unitn.it

---

Abstract

We understand actions from both observation and written text, pointing to a common neural representation of action concepts. However, which parts of the brain encode action concepts independently of stimulus type, say language or visual observation, is an unresolved question in neuroscience. An overlap of activation for action observation and sentence comprehension has been observed in frontoparietal and occipitotemporal cortex, but it is unclear whether this overlap points to the access of stimulus-general action concepts or result from functionally independent neural populations that are activated via one stimulus type but not the other**.** Using fMRI-based MVPA we found that frontoparietal areas encode representations of observed actions and corresponding written sentences in an overlapping way, but these representations did not generalize across stimulus type. By contrast, only left lateral posterior temporal cortex (LPTC) encoded action representations that generalize across observed action scenes and sentences. This observed stimulus generality was not modulated by verbalization or visual imagery. The representational organization of stimulus-general action information in LPTC could be predicted from models that describe basic agent-patient relations (object- and person-directedness) and the general semantic similarity between actions. Person- and object-directedness were topographically organized along a gradient in temporal cortex reminiscent of more posterior object distinctions, pointing to an overarching organizational principle of action and object knowledge in temporal and occipital cortex. Together, our findings revealed fundamentally distinct generalization profiles of temporal and frontoparietal cortex and suggest that LPTC represents aspects of action concepts by integrating more basic precursors in posterior cortex.

Introduction

Our knowledge about things and events in the world is represented at multiple levels, from specific perceptual details (e.g. the movement of a body part) to more general, conceptual aspects (e.g. that a movement serves and is meant to *give* something to someone). Where these levels are represented in the brain is a central issue in neuroscience but remains unresolved (Martin and Chao, 2001; Caramazza and Mahon, 2003; Binder and Desai, 2011; Ralph et al., 2017). While there is considerable progress in understanding the representation

of objects, the representation of action knowledge remains particularly controversial (Rizzolatti and Sinigaglia, 2010; Gallese et al., 2011; Oosterhof et al., 2013; Caramazza et al., 2014). A popular view is that higher-level conceptual aspects of actions are encoded in frontoparietal areas, possibly overlapping with the motor system, whereas perceptual action details such as body parts and movements are encoded in posterior temporal areas in closer proximity to the visual system (Gallese and Lakoff, 2005; Patterson et al., 2007; Rizzolatti and Sinigaglia, 2010). This conception has recently been challenged by demonstrating that posterior temporal cortex encodes action representations (e.g. of opening and closing) that generalize across a range of perceptual features, such as the body parts (Vannuscorps et al., 2018) and movements used to carry out an action (Wurm and Lingnau, 2015; Vannuscorps et al., 2018), the type of object involved in an action (Wurm and Lingnau, 2015; Vannuscorps et al., 2018), and whether an action is recognized from photographs or videos (Hafri et al., 2017). These findings suggest that temporal cortex encodes action representations that abstract away from various details of a perceived action. However, these studies also found that anterior parietal cortex represents actions that generalize across perceptual details (see also Leshinskaya and Caramazza, 2015). Likewise, both areas are also activated during the semantic processing of action words, which lack specific perceptual details of concrete action exemplars (Binder et al., 2009; Watson et al., 2013). These findings raise a puzzling question: If both posterior temporal and anterior parietal cortex are capable of representing actions at similar, high levels of generality, what are their different roles in recognition and memory? It appears unlikely that the two regions have identical functional profiles and store the same, possibly conceptual-level, information in a duplicate way.

Critically, representations of perceptual details are tied to a specific modality, or stimulus type, whereas conceptual representations are generally accessible via different types of stimuli, e.g. via observation or via reading a text. Neuroimaging studies revealed that understanding actions from observation and from written sentences activates overlapping brain networks in prefrontal and parietal cortex as well as in occipitotemporal brain areas, specifically in posterior middle temporal gyrus (pMTG) and surrounding areas in lateral posterior temporal cortex (LTPC) (see also Martin et al., 1995; Aziz-Zadeh et al., 2006; Spunt and Lieberman, 2012). This overlap in activation is usually taken as evidence for the recruitment of the same neural representations accessed during both action observation and sentence comprehension, which would suggest that these representations encode action knowledge at stimulus-independent, conceptual levels. However, overlap in activation is not necessarily due to activation of shared representations (Martin, 2016). Instead, a brain region may house spatially overlapping but functionally independent neural populations that are activated via one stimulus type but not the other. To date, it remains unresolved whether any of the identified brain regions represent conceptual aspects of actions that can be accessed by different kinds of stimuli, such as videos or sentences, a necessary condition of conceptual representation.

Here, we applied a more stringent approach, crossmodal MVPA (Oosterhof et al., 2010; Fairhall and Caramazza, 2013), to identify action representations that are action-specific but at the same time generalize across perception and understanding of visual scenes and sentences. By training a classifier to discriminate actions observed in videos and testing the

same classifier on its accuracy to discriminate corresponding action sentences, this approach targets action representations that are commonly accessed by both stimulus types. In addition, we used a second, conservative criterion to demonstrate conceptual representation: Neural activity patterns associated with different actions should be less or more similar to each other depending whether the actions share fewer or more conceptual features with each other. For example, the action of opening should be more similar to closing as compared to taking, and all three actions should be more similar to each other as compared to communicating actions. Hence, if stimulus-independent representations really encode action concepts then their similarity to each other should not be random but follow semantic principles. We used crossmodal representational similarity analysis (RSA) to test whether semantic models are capable of predicting the similarities of crossmodal action representations, allowing us to investigate the building blocks of action concepts in the brain. We hypothesized that action concepts are represented as propositional structures that consist of more basic components (Schank, 1973), such as agent-patient relations that describe whether or not an action is directed toward other persons or toward objects (Wurm et al., 2017).

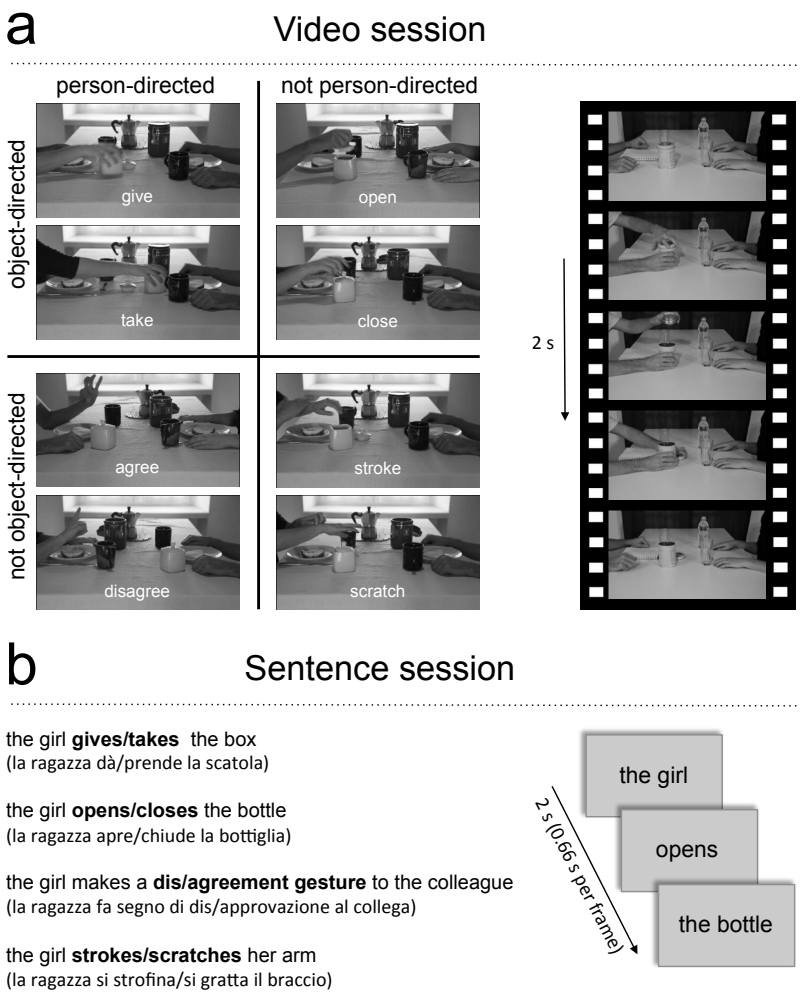


Figure 1. Experimental Design. The video session (A) consisted of action videos (8 actions * 24 video exemplars per action, 2s per video). The sentence session consisted of verbal action descriptions corresponding to the actions shown in the videos (8 actions * 24 sentence exemplars per action, 2s per sentence).

Results

In two fMRI sessions, participants recognized actions presented in videos and corresponding visually-presented sentences (Fig. 1). Participants performed catch trial detection tasks by responding to incomplete or meaningless action videos and grammatically or semantically incorrect sentences. The session order was balanced across participants.

Crossmodal action classification. Using searchlight analysis (Kriegeskorte et al., 2006), we performed to kinds of MVPA to identify brain regions that encode stimulus-general and stimulus-specific action representations. To identify action representations that generalize across stimulus type, we trained a classifier to discriminate actions from video stimuli and tested the classifier on the sentences, and vice versa. This analysis identified a single cluster in the left LPTC, peaking in pMTG and extending dorsally into posterior superior temporal sulcus (pSTS) and ventrally into inferior temporal gyrus (Fig. 2A, Supplementary Table 2). By contrast, if classifiers were trained and tested within action videos or within sentences, actions could be discriminated in more extended networks overlapping in occipitotemporal and frontoparietal areas (Fig. 2B), in line with previous findings (Aziz-Zadeh et al., 2006; Spunt and Lieberman, 2012). Overall, classification accuracies were higher for videos than for sentences. Apart from these general activation differences, some areas appeared to be particularly sensitive to observed actions (left and right lateral occipitotemporal cortex and anterior inferior parietal cortex), whereas other areas were particularly sensitive to sentences (left posterior superior temporal gyrus and inferior frontal cortex, anterior temporal lobes). Critically, frontoparietal areas discriminated both action videos and sentences, but the absence of crossmodal decoding (see Supplementary Fig. 1 and Supplementary Table 3 for Bayesian model comparisons) suggests that these representations do not generalize across stimulus type.

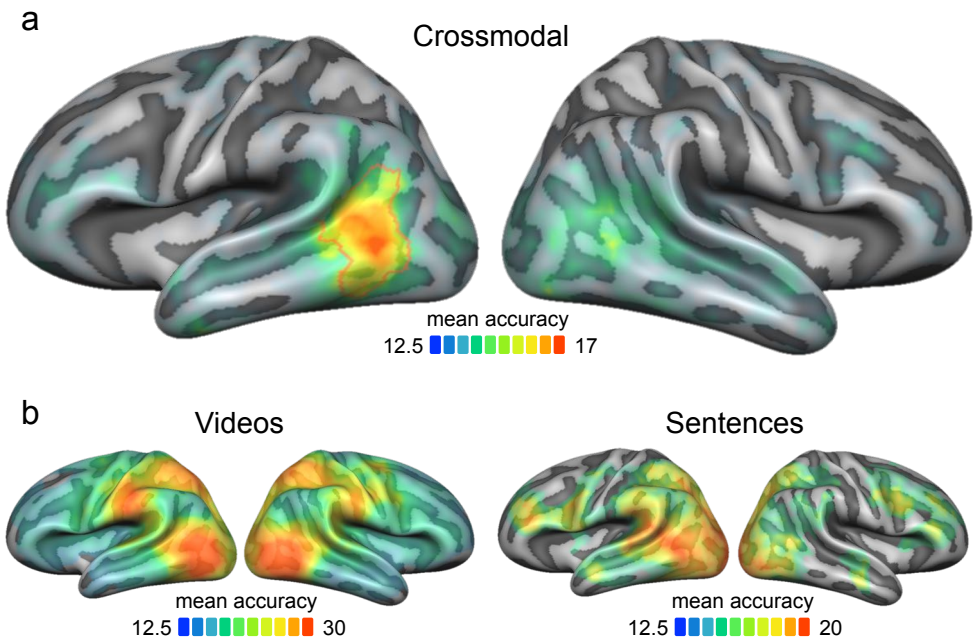


Figure 2. Multiclass action decoding searchlight (mean accuracies, chance = 12.5%). (A) Crossmodal classification. Red outlines indicate TFCE-corrected areas. (B) Within-modality classification. Maps thresholded using TFCE correction.

Could the generalization across stimulus type in left LPTC be explained by verbalization or visual imagery? Our experimental design allowed testing these possibilities: As we balanced the order of video and sentence sessions across participants, we would expect stronger verbalization in the participant group that started with the sentence session, and stronger imagery in the participant group that started with the video session. Contrasting the decoding maps of the two groups, however, revealed no significant differences in left LPTC ($t(20) = 0.04$, $p = 0.97$, two-tailed; Supplementary Fig. 3A). In addition, we found no significant correlations between decoding accuracies in LPTC and scores obtained in a post-fMRI rating on verbalization ($r(19) = 0.26$, $p = 0.13$, one-tailed) and visual imagery ($r(19) = -0.31$, $p = 0.97$, one-tailed; Supplementary Fig. 3B and C). Thus, visual imagery and verbalization seem to be unlikely to account for the observed crossmodal decoding effects, suggesting that action representations in LPTC could be accessed via both stimulus types independently of each other.

Crossmodal multiple regression RSA. The accessibility of action-specific representations during both action observation and sentence reading suggests a central role of LPTC in action understanding. What exactly do these representations encode? Using multiple regression RSA, we analyzed the structure of action representations in LPTC in more detail. To this end, we extracted the pairwise crossmodal classification accuracies to construct a neural dissimilarity matrix, which reflects how well the actions could be discriminated from each other, and thus how dissimilar the action representations are to each other (Fig. 3A). We found that the neural action dissimilarity in LPTC could be predicted by models of person- and object-directedness, which were based on post-fMRI ratings of how much the actions shown in the experiment took into account the reactions of other persons (person-directedness) and how much the actions involved an interaction with physical, nonliving objects (object-directedness, Fig. 3B). Additional variance in the neural representational similarity could be explained by a more general model of semantic action similarity that group actions based on a combination of semantic action relations (Miller et al., 1990). Other models that were included in the regression to control for factors of no interest (semantic object similarity, task difficulty, unusualness of actions) could not explain further variance. RSA effects were robust across ROI size and number of models included in the analysis (Supplementary Figure 4). In addition, we found that the rating-based models of person- and object-directedness correlated significantly better with the neural action dissimilarities than simple categorical models that were based on the main stimulus dimensions used in the experiment (Fig. 3C; person-directedness: person-directedness: $Z = 3.12$, $p = 0.001$, object-relatedness: $Z = 2.62$, $p = 0.009$; two-tailed signed-rank test). This suggests that the neural representational organization in LPTC indeed reflect stimulus variations in person- and object-directedness as measured by behavioral judgments rather than other hidden factors that may have accidentally covaried with the stimulus dimensions.

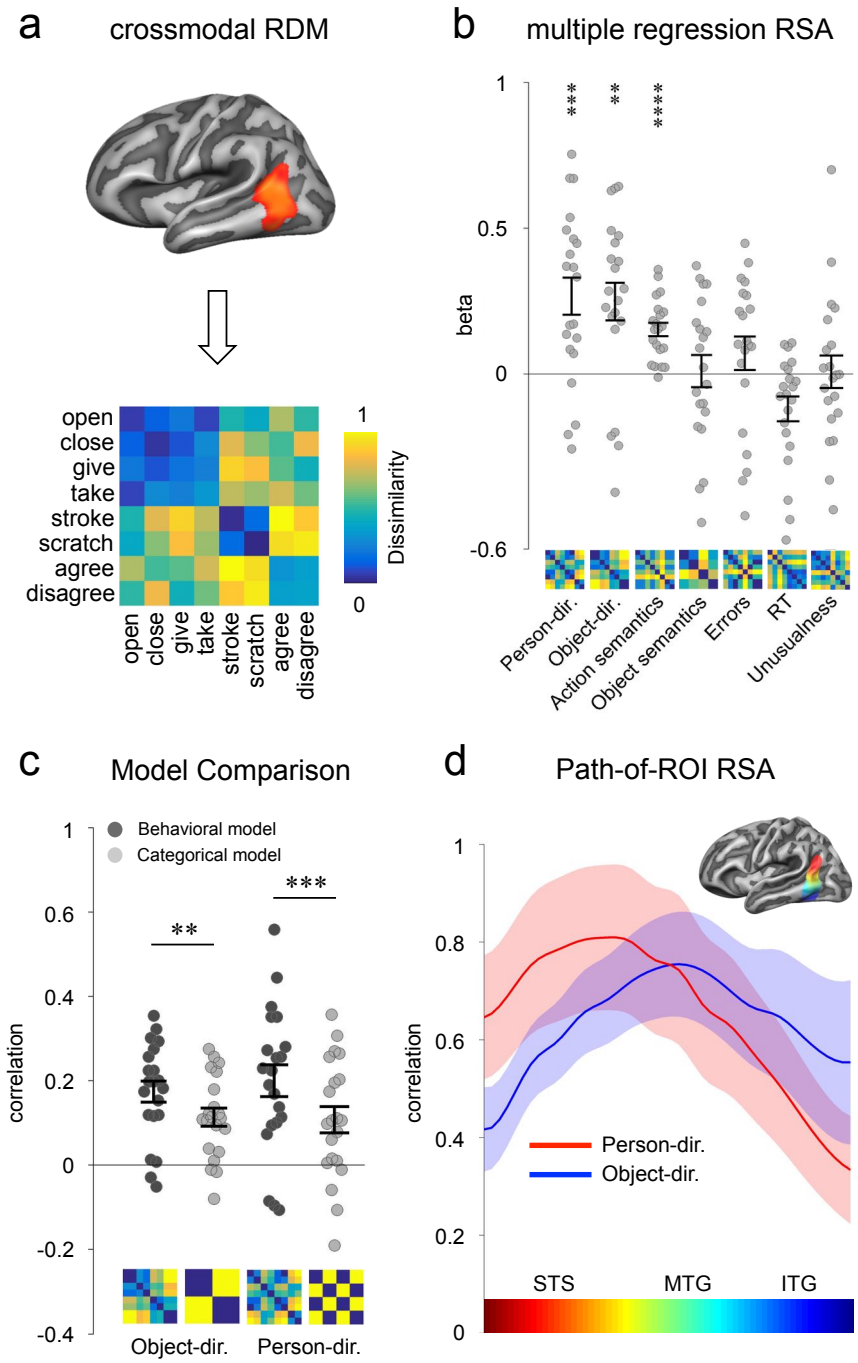


Figure 3. Crossmodal RSA. (A) Classification matrices were extracted from left LPTC, averaged, and converted into a dissimilarity matrix for each participant (displayed matrix averaged across participants). (B) Multiple regression RSA. Person- and object-directedness were based on ratings; WordNet action and object models were based on taxonomical distance (path length); crossmodal control models were based on behavioral experiments (Errors, Reaction Times) and ratings for unusualness of actions presented in videos and sentences. (C) Comparison between performance of behavioral (rating-based) and categorical models of person- and object-directedness using correlation-based RSA. (D) Path-of-ROI analysis. Mean RSA correlation coefficients plotted as a function of position along dorsal-ventral axis. X-axis color bar corresponds to ROI colors. Asterisks in B and C indicate FDR-corrected effects: ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$. Error bars/lines indicate SEM.

Path-of-ROI RSA. Crossmodal action information could be decoded from a relatively large cluster in LPTC spanning from posterior superior temporal sulcus to inferior temporal gyrus. Is this area parceled into distinct subregions specialized for certain aspects of actions? Based on previous findings (Wurm et al., 2017), we hypothesized that conceptual information related to person- and object directedness is not distributed uniformly in LPTC but rather follow a distinctive functional topography that parallels known gradients in more posterior areas segregating animate and inanimate object knowledge (Chao et al., 1999; Konkle and Caramazza, 2013) and biological and functional tool motion (Beauchamp et al., 2002). To investigate how the representational content changes from dorsal to ventral temporal cortex, we plotted the performance of the person-directedness and object-relatedness models as a function of position along a cortical vector from the dorsal to the ventral end of the crossmodal decoding cluster (Figure 3D). In line with our prediction, we found that the person-directedness model explained most variance in dorsal LOTC at the level of pSTS (Talairach z = 9) whereas the object- directedness model peaked more ventrally (Talairach z = 3) at the level of pMTG (ANOVA interaction POSITION x MODEL: $F(15,300) = 4.23$, $p < 0.001$).

Discussion

Conceptual representations should generally be accessible independently of the modality or type of stimulus, e.g., whether one observes an action or reads a corresponding verbal description. Here we show that overlap of activation alone during action observation and sentence comprehension cannot be taken as evidence for the access of stimulus-independent, conceptual representations. Using crossmodal MVPA, we found that only the left LPTC encodes information that is both action-specific and at the same time generalizes across action scenes and sentences. Crossmodal classification could not be explained by tendencies to verbalize or to visually imagine the actions, making it unlikely that the decoded information in LPTC was in a verbal format, triggered by both the sentences and by verbalizations of the action scenes, or in a visual format, triggered by the observed action scenes and imagery of the sentences. Moreover, crossmodal RSA revealed that the organization on stimulus-general information in LPTC could be predicted by models that describe basic agent-patient relations – person- and object-directedness – as well as more complex semantic relations between action concepts such as whether actions are subparts of or oppose each other. Together, these findings provide clear evidence that left LPTC encodes conceptual aspects of actions that can be accessed by different stimulus types like action scenes and sentences. By contrast, frontoparietal areas discriminated both action scenes and sentences, but the decoded representations did not generalize across the two stimulus types. Representations in LPTC thus seem to be more general and abstract as compared to frontoparietal representations, which appear to capture more specific details and properties of the different stimulus types. Notably, observed actions are more specific and rich in details compared to sentences. The stronger decoding of action scenes relative to sentences appears to reflect this difference in richness of detail. The decoded frontoparietal representations might capture information about how specifically an action is carried out. Such motor-related

representations might be triggered by action scenes reflecting specific aspects of the action (Negri et al., 2007; Buxbaum and Kalenine, 2010), such as the kinematics of an action or the particular grip used on an object, whereas the motor-related representations triggered by sentences would be less specific, more variable, and less robust. Following this view, frontoparietal motor-related representations are activated following conceptual activation in LPTC, in line with the finding that TMS-induced perturbation of pMTG not only impedes semantic processing of action verbs but also disrupts increased motor excitability for motor as compared to non-motor verbs (Papeo et al., 2015). Likewise, dysplasic individuals born without arms recognize hand actions, for which they do not have corresponding motor representations, as fast and as accurately as typically developed participants suggesting that motor-related representations are not necessary in accessing action concepts (Vannuscorps and Caramazza, 2016). Notably, neural activity and representational content in frontoparietal cortex does not differ between dysplasic and typically developed individuals during hand action observation (Vannuscorps et al., 2018). An additional possibility is therefore that frontoparietal areas are involved in processing non-motor information related to the stimuli, for example in the service of anticipating how action scenes and sentences unfold (Schubotz, 2007; Kilner, 2011; Willems et al., 2016). In support of this view, premotor and parietal areas have been shown to be engaged in the prediction of dynamic stimuli of different types and modalities (Schubotz and von Cramon, 2003), even if stimuli lack any motor-relevant pragmatic properties (Schubotz and Yves von Cramon, 2002).

Whereas the crossmodal decoding analyses reported here focused on identifying the neural substrate of conceptual action representations, the crossmodal RSA allowed us to investigate the structure of these representations in more detail. We found that the similarity of neural patterns associated with the actions tested in this experiment could be predicted by models that describe whether actions are directed toward other persons or toward inanimate objects as well as a more general model of semantic action similarity. Representations sensitive to person- and object-directedness followed distinctive functional topographies: Neural populations in the dorsal part of LPTC were more sensitive to detect whether an action is person-directed or not, whereas neural populations in more ventral parts were more sensitive to detect whether an action is object-directed or not. This result replicates the previous observation of a dorsal-ventral gradient for observed actions (Wurm et al., 2017) and demonstrates that a similar (but left-lateralized) gradient exists also for stimulus-general action representations. The topographical distinction of person- and object-directedness resembles related gradients that are typically found in adjacent posterior/ventral areas: In lateral occipitotemporal cortex, overlapping with the posterior part of the cluster found in the present study, dorsal subregions are preferentially activated by animate objects (e.g. animals and body parts) whereas ventral subregions are preferentially activated by inanimate objects (e.g. manipulable artifacts and tools) (Chao et al., 1999; Konkle and Caramazza, 2013). Likewise, dorsal subregions are preferentially activated by body movements whereas ventral subregions are preferentially activated by action-specific tool movements (Beauchamp et al., 2002). Here we demonstrate a continuation of this distinction for more complex action components that cannot be explained by visual stimulus properties. The topographic alignment of object, object motion, and modality-general action representation points to an

overarching organizational principle of action and object knowledge in temporal and occipital cortex: Knowledge related to persons is represented in dorsal posterior occipitotemporal cortex and specifically associated with (and form the perceptual basis of) biological motion and person-directed actions in dorsal LPTC. Furthermore, knowledge about inanimate manipulable objects is represented in ventral occipitotemporal cortex and specifically associated with action-related object motion patterns and object-directed actions in ventral LPTC. The functionally parallel organization of object and action knowledge in lateral occipital and temporal cortex, respectively, suggest the important role of domain-specific object-action connections along the posterior-anterior axis, in agreement with the view that connectivity plays a fundamental role in shaping the functional organization of distinct knowledge categories (Mahon and Caramazza, 2011; Osher et al., 2016).

In conclusion, the different generalization profiles of temporal and frontoparietal areas point to a hierarchy of action representation from specific perceptual stimulus features in occipitotemporal areas to more general, conceptual aspects in left LPTC and back to stimulus-specific representations in frontoparietal cortex. We propose that the topographic organization of conceptual action knowledge is (at least partially) determined by the representation of more basic precursors, such as animate and inanimate entities.

Materials and Methods

Participants. Twenty-one right-handed native Italian speakers (9 females; mean age, 23.8 years; age range, 20-36 years) participated in this experiment. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disease. One subject was excluded due to poor behavioral performance in the task (accuracy two standard deviations below the group mean). All procedures were approved by the Ethics Committee for research involving human participants at the University of Trento, Italy.

Stimuli. The video stimuli consisted of 24 exemplars of eight hand actions (192 action videos in total) as used in Wurm et al. (2017). The actions varied along two dimensions, person-directedness (here defined as the degree to which actions take into account the actions and reactions of others) and object-directedness (here defined as the degree to which actions involve the interaction with physical inanimate objects), resulting in four action categories: change of possession (object-directed/person-directed): *give*, *take*; object manipulation (object-directed/nonsocial): *open*, *close*; communication (not object-directed/person-directed): *agree*, *disagree*; body/contact action (not object-directed/not person-directed): *stroke*, *scratch*. By using 24 different exemplars for each action we increased the perceptual variance of the stimuli to ensure that classification is trained on conceptual rather than perceptual features. Variance was induced by using two different contexts, three perspectives, two actors, and six different objects that were present or involved in the actions (kitchen context: sugar cup, honey jar, coffee jar; office context: bottle, pen box, aluminum box). Video catch trials consisted of six deviant exemplars of each of the eight actions (e.g., meaningless gestures or object manipulations, incomplete actions; 48 catch trial videos in total). All 240 videos were identical in terms of action timing, i.e., the videos started with

hands on the table, followed by the action, and ended with hands moving to the same position of the table. Videos were gray scale, had a length of 2 s (30 frames per second), and a resolution of 400 x 225 pixels.

The sentence stimuli were matched with the video stimuli in terms of stimulus variance (24 sentences of eight actions; 192 sentence videos in total). All sentences had the structure subject-verb-object. For each action, we first defined the corresponding verb phrase: *dà* (*s/he gives*), *prende* (*s/he takes*), *apre* (*s/he opens*), *chiude* (*s/he closes*), *si strofina* (*s/he rubs her/his*), *si gratta* (*s/he scratches her/his*), *fa segno di approvazione a* (*s/he makes a sign of agreement to*), *fa segno di disapprovazione a* (*s/he makes a sign of disagreement to*). To create 24 exemplars per action, we crossed each verb phrase with 6 subjects: *lei, lui, la ragazza, il ragazzo, la donna, l'uomo* (*she, he, the girl, the boy, the woman, the man*) and with 4 objects, which matched the objects used in the videos as much as possible. Change of possession: *la scatola, il vaso, il caffe, lo zucchero* (*the box, the jar, the coffee, the sugar*); object manipulation: *la bottiglia, il barattolo, la cassetta, l'astuccio* (*the bottle, the can, the casket, the pencil case*); communication: *l'amica, l'amico, il college, la collega* (*friend, colleague*); body action: *il braccio, la mano, il gomito, l'avambraccio* (*the arm, the hand, the elbow, the forearm*). As the crossmodal analysis focuses on the generalization across sentence and video stimuli, perceptual and syntactic differences between action sentences, such as sentence length and occurrence of prepositions, were ignored. Catch trial sentences consisted of six grammatically incorrect or semantically odd exemplars of each of the eight actions (e.g., *lei apre alla bottiglia* (*she opens to the bottle*), *lui da l'amica* (*he gives the friend*); 48 catch trial sentences in total). The sentences were presented superimposed on light grey background (400 x 225 pixels) in three consecutive chunks (subject, verb phrase, object), with each chunk shown for 666 ms (2 s per sentence), using different font types (Arial, Times New Roman, Comic Sans MS, MV Boli, MS UI Gothic, Calibri Light) and font sizes (17-22) to increase the perceptual variance of the sentence stimuli (balanced across conditions within experimental runs).

In the scanner, stimuli were back-projected onto a screen (60 Hz frame rate, 1024 x 768 pixels screen resolution) via a liquid crystal projector (OC EMP 7900, Epson Nagano, Japan) and viewed through a mirror mounted on the head coil (video presentation 6.9° x 3.9° visual angle). Stimulus presentation, response collection, and synchronization with the scanner were controlled with ASF (Schwarzbach, 2011) and the Matlab Psychtoolbox-3 for Windows (Brainard, 1997).

Experimental Design. For both video and sentence sessions, stimuli were presented in a mixed event-related design. In each trial, videos/sentences (2 s) were followed by a 1 s fixation period. 18 trials were shown per block. Each of the nine conditions (eight action conditions plus one catch trial condition) was presented twice per block. Six blocks were presented per run, separated by 10 s fixation periods. Each run started with a 10 s fixation period and ended with a 16 s fixation period. In each run, the order of conditions was first-order counterbalanced (Aguirre, 2007). Each participant was scanned in two sessions (video and sentence session), each consisting of 4 functional scans, and one anatomical scan. The order of sessions was counterbalanced across participants (odd IDs: videos-sentences, even

IDs: sentences-videos). For each of the nine conditions per session there was a total of 2 (trials per block) x 6 (blocks per run) x 4 (runs per session) = 48 trials per condition. Each of the 24 exemplars per action condition was presented twice in the experiment.

Task. Before fMRI, we instructed and trained participants for the first session only (videos or sentences). The second session was instructed and practiced within the scanner after the four runs of the first session. Participants were asked to attentively watch the videos [read the sentences] and to press a button with the right index finger on a response button box whenever an observed action was meaningless or performed incompletely or incorrectly [whenever a sentence was meaningless or grammatically incorrect]. The task thus induced the participants to understand the actions, while minimizing the possibility of additional cognitive processes that might be different between the actions but similar across sessions. For example, tasks that require judgments about the actions along certain dimensions such as action familiarity (Quandt et al., 2017) might lead to differential neural activity related to the preparation of different responses, which could be decodable across stimulus type. Participants could respond either during the video/sentence or during the fixation phase after the video/sentence. To ensure that participants followed the instructions correctly, they completed a practice run before the respective session. Participants were not informed about the purpose and design of the study before the experiment.

Post fMRI survey. After the fMRI session, participants judged the degree of person- and object directedness of the actions in the experiment. For each action, participants answered ratings to the following questions: *Object-directedness*: "How much does the action involve the interaction with a physical, inanimate object?" *Person-directedness*: "How much does the action take into account the actions and reactions of another person?" In addition, they were asked to estimate how much they verbalized the actions during the video session (*Verbalization*: "During watching the action videos, did you verbalize the actions, that is, did you have verbal descriptions (words, sentences) in your mind as if you were talking to yourself?"), how much they visually imagined the action in the sentence session (*Imagery*: "During reading the sentences, did you visually imagine concrete action scenes?"), and how similar their verbal descriptions were to the sentences (*Verbalization-sentence correspondence*) and how similar the imagined action scenes were to the videos; (*Imagery-video correspondence*, respectively). For all ratings, 6-point Likert scales (from 1 = not at all to 6 = very much) were used.

Representational dissimilarity models (RDMs). To investigate the representational organization of voxel patterns that encode crossmodal action information, we tested the following model RDMs:

To generate models of *Person- and Object-directedness,* we computed pairwise Euclidean distances between the group-averaged responses to each of the actions from the ratings for person- and object directedness. For comparison, we also tested categorical models that segregated the actions along person- and object directedness without taking into account more subtle action-specific variations.

To generate a model of semantic relationship between the actions (*Action semantics* hereafter) that is not solely based on either person- or object-directedness, but rather reflect

semantic relations between action concepts, we computed hierarchical distances between action concepts based on WordNet 2.1 (Miller, 1995). This model captures a combination of semantic relations between actions, such as whether actions are subparts of each other (e.g. drinking entails swallowing) or oppose each other (e.g. opening and closing). We used WordNet because it is supposed to reflect conceptual-semantic rather than syntactic relations between words. Semantic relations should therefore be applicable to both action sentences and videos. We selected action verbs by identifying the cognitive synonyms (synsets) in Italian that matched the verbs used in the action sentences and that corresponded best to the meaning of the actions ('open.v.01', 'close.v.01', 'give.v.03', 'take.v.08', 'stroke.v.01', 'scratch.v.03', 'agree.v.01', 'disagree.v.01'). We computed pairwise semantic distances between the eight actions using the shortest path length measure, which reflects the taxonomical distance between action concepts.

In a similar way, we generated a model of semantic relationship between the target objects (inanimate objects, body parts, persons) of the actions in the 4 action categories (*Object semantics*). As for the action verbs, we selected Italian synsets that matched the object nouns in the sentences. As there were four objects per action category, the distances were averaged within each category.

To generate models of task difficulty (*RT* and *Errors*), an independent group of participants (N=12) performed a behavioral 2-alternative forced choice experiment that had the same design and instruction as the fMRI experiment except that participants responded with the right index finger to correct action videos/sentences (action trials) and with the right middle finger to incorrect action videos/sentences (catch trials). Errors and RTs of correct responses to action trials were averaged across participants. The Error RDM was constructed by computing the pairwise Euclidean distances between the 8 accuracies of the sentence session and the 8 accuracies of the video session. The model thus reflects how similar the eight actions are in terms of errors made across the two sessions. The RT RDM was constructed in a similar way except that the RTs of each session were zscored before computing the distances to eliminate session-related differences between video and sentence RTs.

To generate a model reflecting potential differences in saliency due to *Unusualness* between the actions, we asked the participants of the behavioral experiment described in the previous paragraph to judge how unusual the actions were in the videos and sentences, respectively (using 6-point Likert scales from 1 = not at all to 6 = very much). The Unusualness RDM was constructed by computing the pairwise Euclidean distances between the 8 mean responses to the sentence session and the 8 mean responses to the video session.

Data acquisition. Functional and structural data were collected using a 4 T Bruker MedSpec Biospin MR scanner and an 8-channel birdcage head coil. Functional images were acquired with a T2*-weighted gradient echo-planar imaging (EPI) sequence with fat suppression. Acquisition parameters were a repetition time of 2.2 s, an echo time of 33 ms, a flip angle of 75°, a field of view of 192 mm, a matrix size of 64 x 64, and a voxel resolution of 3 x 3 x 3 mm. We used 31 slices, acquired in ascending interleaved order, with a thickness of 3 mm and 15 % gap (0.45 mm). Slices were tilted to run parallel to the superior temporal sulcus. In each functional run, 176 images were acquired. Before each run we performed an additional

scan to measure the point-spread function (PSF) of the acquired sequence to correct the distortion expected with high-field imaging (Zaitsev et al., 2004).

Structural T1-weigthed images were acquired with an MPRAGE sequence (176 sagittal slices, TR = 2.7 s, inversion time = 1020 ms, FA = 7°, 256 x 224 mm FOV, 1 x 1 x 1 mm resolution).

Preprocessing. Data were analyzed using BrainVoyager QX 2.8 (BrainInnovation) in combination with the BVQXTools and NeuroElf Toolboxes and custom software written in Matlab (MathWorks). Distortions in geometry and intensity in the echo-planar images were corrected on the basis of the PSF data acquired before each EPI scan (Zeng and Constable, 2002). The first 4 volumes were removed to avoid T1 saturation. The first volume of the first run was aligned to the high-resolution anatomy (6 parameters). Data were 3D motion corrected (trilinear interpolation, with the first volume of the first run of each participant as reference), followed by slice time correction and high-pass filtering (cutoff frequency of 3 cycles per run). Spatial smoothing was applied with a Gaussian kernel of 8 mm FWHM for univariate analysis and 3 mm FWHM for MVPA. Anatomical and functional data were transformed into Talairach space using trilinear interpolation.

Action classification. For each participant and run, a general linear model (GLM) was computed using design matrices containing 16 action predictors (2 predictors per action, each based on 6 trials selected from the first half (blocks 1-3) or the second half (blocks 4-6) of each run), catch trials, and of the 6 parameters resulting from 3D motion correction (x, y, z translation and rotation). Each predictor was convolved with a dual-gamma hemodynamic impulse response function (Friston et al., 1998). Each trial was modeled as an epoch lasting from video/sentence onset to offset (2 s). The resulting reference time courses were used to fit the signal time courses of each voxel. In total, this procedure resulted in 8 beta maps per action condition and session.

Searchlight classification (Kriegeskorte et al., 2006) was performed for each participant separately in volume space using searchlight spheres with a radius of 12 mm and a linear support vector machine (SVM) classifier as implemented by the CoSMoMVPA toolbox (Oosterhof et al., 2016) and LIBSVM (Chang and Lin, 2011). We demeaned the data for each multivoxel beta pattern in a searchlight sphere across voxels by subtracting the mean beta of the sphere from each beta of the individual voxels. In all classification analyses, each action was discriminated from each of the remaining seven actions in a pairwise manner ("one-against-one" multiclass decoding). For searchlight analyses, accuracies were averaged across the eight actions (accuracy at chance = 12.5%) and assigned to the central voxel of each searchlight sphere. In the crossmodal action classification, we trained a classifier to discriminate the actions using the data of the video session and tested the classifier on its accuracy at discriminating the actions using the data of the sentence session. The same was done vice versa (training on sentence data, testing on video data), and the resulting accuracies were averaged across classification directions. In the within-video action classification, we decoded the eight actions of the video session using leave-one-beta-out cross validation: We trained a classifier to discriminate the actions using 7 out of the 8 beta patterns per action. Then we tested the classifier on its accuracy at discriminating the actions using the held out

data. This procedure was carried out in 8 iterations, using all possible combinations of training and test patterns. The resulting classification accuracies were averaged across the 8 iterations. The same procedure was used in the within-sentence action classification. Individual accuracy maps were entered into a one-sample t-test to identify voxels in which classification was significantly above chance. Statistical maps were thresholded using Threshold-Free Cluster Enhancement (TFCE; Smith and Nichols, 2009) as implemented in the CoSMoMVPA Toolbox (Oosterhof et al., 2016). We used 10000 Monte Carlo simulations and a one-tailed corrected cluster threshold of $p = 0.05$ ($z = 1.65$). Maps were projected on a cortex-based aligned group surface for visualization. Significant decoding accuracies below chance (using both two and one-tailed tests) were not observed.

Bayes factor maps were computed using R (version 3.4.2) and the BayesFactor package (version 0.9.12) (Morey et al., 2015). For each voxel, we computed directional Bayes factors to compare the hypotheses that the standardized effect is at chance (12.5%) vs. above chance, using a default Cauchy prior width of $r = 0.707$ for effect size.

Crossmodal representational similarity analysis (RSA). For further investigation of the representational organization in the voxels that encode crossmodal action information we performed an ROI-based multiple regression RSA (Kriegeskorte et al., 2008; Bracci et al., 2015) using crossmodal classification accuracies (Fairhall and Caramazza, 2013). ROIs were defined based on TFCE-corrected clusters identified in the classification analysis. From each ROI voxel, we extracted the pairwise classifications of the crossmodal action decoding, resulting in voxelwise 8 x 8 classification matrices. Notably the selection of action-discriminative voxels, which is based on on-diagonal data of the pairwise classification matrix, does not bias toward certain representational organizations investigated in the RSA, which is based on off-diagonal data of the pairwise classification matrix (see below). Classification matrices were inverted (100 – accuracy) and averaged across voxels, resulting in one representational dissimilarity matrix (RDM) per ROI and participant.

The off-diagonals (i.e., the lower triangular parts) of the neural and the model RDMs were vectorized, z-scored, and entered as independent and dependent variables, respectively, into a multiple regression. We tested for putative collinearity between the models by computing condition indices (CI), variance inflation factors (VIF), and variance decomposition proportions (VDP) using the colldiag function for Matlab. The results of these tests (max. CI=3, max. VIF=2.6, max DVP=0.8) revealed no indications of potential estimation problems (Belsley et al., 1980). Correlation coefficients were Fisher transformed and entered into one-tailed signed-rank tests. Results were FDR-corrected for the number of models included in the regression.

To compare the performance of individual models, we used correlation-based RSA, i.e., we computed rank correlations between the off-diagonals of neural and model RDMs using Kendall's τ_A as implemented by the toolbox for representational similarity analysis (Nili et al., 2014). Comparisons between rank correlations were computed using FDR-corrected paired two-tailed signed-rank tests.

To investigate the performance of models along the dorsal-ventral axis, we used a path-of-ROI analysis (Konkle and Caramazza, 2013) as described in Wurm et al. (2017): Dorsal and

ventral anchor points were based on the most dorsal and ventral voxels, respectively, of the cluster identified in the crossmodal action classification (Talairach x/y/z; dorsal: -43/-58/20; ventral: -43/-50/-10). Anchor points were connected with a straight vector on the flattened surface. Along this vector, a series of partially overlapping ROIs (12 mm radius, centers spaced 3 mm) was defined. From each ROI, neural RDMs were extracted, averaged, and entered into correlation-based RSA as described above. Resulting correlation coefficients were averaged across participants and plotted as a function of position on the dorsal-ventral axis.

Data sharing. Materials, neuroimaging data, and code are available upon request.

Acknowledgements

This work was supported by the Provincia Autonoma di Trento and the Fondazione CARITRO (SMC). We thank Valentina Brentari for assistance in preparing the verbal stimulus material and with data acquisition.

References

Aguirre GK (2007) Continuous carry-over designs for fMRI. Neuroimage 35:1480-1494.

Aziz-Zadeh L, Wilson SM, Rizzolatti G, Iacoboni M (2006) Congruent embodied representations for visually presented actions and linguistic phrases describing actions. Curr Biol 16:1818-1823.

Beauchamp MS, Lee KE, Haxby JV, Martin A (2002) Parallel visual motion processing streams for manipulable objects and human movements. Neuron 34:149-159.

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: Identifying influential data and sources of collinearity: John Wiley & Sons.

Binder JR, Desai RH (2011) The neurobiology of semantic memory. Trends Cogn Sci 15:527-536.

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767-2796.

Bracci S, Caramazza A, Peelen MV (2015) Representational Similarity of Body Parts in Human Occipitotemporal Cortex. J Neurosci 35:12977-12985.

Brainard DH (1997) The Psychophysics Toolbox. Spatial vision 10:433-436.

Buxbaum LJ, Kalenine S (2010) Action knowledge, visuomotor activation, and embodiment in the two action systems. Ann N Y Acad Sci 1191:201-218.

Caramazza A, Mahon BZ (2003) The organization of conceptual knowledge: the evidence from category-specific semantic deficits. Trends Cogn Sci 7:354-361.

Caramazza A, Anzellotti S, Strnad L, Lingnau A (2014) Embodied cognition and mirror neurons: a critical assessment. Annu Rev Neurosci 37:1-15.

Chang C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:27.

Chao LL, Haxby JV, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat Neurosci 2:913-919.

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552-10558.

Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R (1998) Event-related fMRI: characterizing differential responses. Neuroimage 7:30-40.

Gallese V, Lakoff G (2005) The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. Cognitive neuropsychology 22:455-479.

Gallese V, Gernsbacher MA, Heyes C, Hickok G, Iacoboni M (2011) Mirror Neuron Forum. Perspectives on Psychological Science 6:369-407.

Hafri A, Trueswell JC, Epstein RA (2017) Neural Representations of Observed Actions Generalize across Static and Dynamic Visual Input. J Neurosci 37:3056-3071.

Jeffreys H (1998) The theory of probability: OUP Oxford.

Kilner JM (2011) More than one pathway to action understanding. Trends Cogn Sci 15:352-357.

Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. J Neurosci 33:10235-10242.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863-3868.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2:4.

Leshinskaya A, Caramazza A (2015) Abstract categories of functions in anterior parietal lobe. Neuropsychologia 76:27-40.

Mahon BZ, Caramazza A (2011) What drives the organization of object knowledge in the brain? Trends Cogn Sci 15:97-103.

Martin A (2016) GRAPES-Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. Psychonomic bulletin & review 23:979-990.

Martin A, Chao LL (2001) Semantic memory and the brain: structure and processes. Curr Opin Neurobiol 11:194-201.

Martin A, Haxby JV, Lalonde FM, Wiggs CL, Ungerleider LG (1995) Discrete cortical regions associated with knowledge of color and knowledge of action. Science 270:102-105.

Miller GA (1995) WordNet: a lexical database for English. Communications of the ACM 38:39-41.

Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to WordNet: An on-line lexical database. International Journal of Lexicography 3:235-244.

Morey RD, Rouder JN, Jamil T, Morey MRD (2015) Package 'BayesFactor'. URL〈 http://cran r-project org/web/packages/BayesFactor/BayesFactor pdf〉(accessed 1006 15).

Negri GA, Rumiati RI, Zadini A, Ukmar M, Mahon BZ, Caramazza A (2007) What is the role of motor simulation in action and object recognition? Evidence from apraxia. Cognitive neuropsychology 24:795-816.

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653-660.

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.

Oosterhof NN, Tipper SP, Downing PE (2013) Crossmodal and action-specific: neuroimaging the human mirror neuron system. Trends Cogn Sci 17:311-318.

Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front Neuroinform 10:27.

Oosterhof NN, Wiggett AJ, Diedrichsen J, Tipper SP, Downing PE (2010) Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. J Neurophysiol 104:1077-1089.

Osher DE, Saxe RR, Koldewyn K, Gabrieli JD, Kanwisher N, Saygin ZM (2016) Structural Connectivity Fingerprints Predict Cortical Selectivity for Multiple Visual Categories across Cortex. Cereb Cortex 26:1668-1683.

Papeo L, Lingnau A, Agosta S, Pascual-Leone A, Battelli L, Caramazza A (2015) The origin of word-related motor activity. Cereb Cortex 25:1668-1675.

Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci 8:976-987.

Quandt LC, Lee YS, Chatterjee A (2017) Neural bases of action abstraction. Biol Psychol 129:314-323.

Ralph MA, Jefferies E, Patterson K, Rogers TT (2017) The neural and computational bases of semantic cognition. Nat Rev Neurosci 18:42-55.

Rizzolatti G, Sinigaglia C (2010) The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat Rev Neurosci 11:264-274.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin & review 16:225-237.

Schank RC (1973) The fourteen primitive actions and their inferences. Memo AIM-183, Stanford Artificial Intelligence Laboratory.

Schubotz RI (2007) Prediction of external events with our motor system: towards a new framework. Trends Cogn Sci 11:211-218.

Schubotz RI, Yves von Cramon D (2002) Dynamic patterns make the premotor cortex interested in objects: influence of stimulus and task revealed by fMRI. Brain Res Cogn Brain Res 14:357-369.

Schubotz RI, von Cramon DY (2003) Functional-anatomical concepts of human premotor cortex: evidence from fMRI and PET studies. Neuroimage 20 Suppl 1:S120-131.

Schwarzbach J (2011) A simple framework (ASF) for behavioral and neuroimaging experiments based on the psychophysics toolbox for MATLAB. Behavior research methods 43:1194-1201.

Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83-98.

Spunt RP, Lieberman MD (2012) Dissociating modality-specific and supramodal neural systems for action understanding. J Neurosci 32:3575-3583.

Vannuscorps G, Caramazza A (2016) Typical action perception and interpretation without motor simulation. Proc Natl Acad Sci U S A 113:86-91.

Vannuscorps G, Wurm M, Striem-Amit E, Caramazza A (2018) Large-scale organization of the hand action observation network in individuals born without hands. bioRxiv:305888.

Watson CE, Cardillo ER, Ianni GR, Chatterjee A (2013) Action concepts in the brain: an activation likelihood estimation meta-analysis. J Cogn Neurosci 25:1191-1205.

Willems RM, Frank SL, Nijhof AD, Hagoort P, van den Bosch A (2016) Prediction During Natural Language Comprehension. Cereb Cortex 26:2506-2516.

Wurm MF, Lingnau A (2015) Decoding actions at different levels of abstraction. J Neurosci 35:7727-7735.

Wurm MF, Caramazza A, Lingnau A (2017) Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. J Neurosci 37:562-575.

Zaitsev M, Hennig J, Speck O (2004) Point spread function mapping with parallel imaging techniques and high acceleration factors: fast, robust, and flexible method for echo-planar imaging distortion correction. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 52:1156-1166.

Zeng H, Constable RT (2002) Image distortion correction in EPI: comparison of field mapping with point spread function mapping. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 48:137-146.

Supporting Information

SI Results

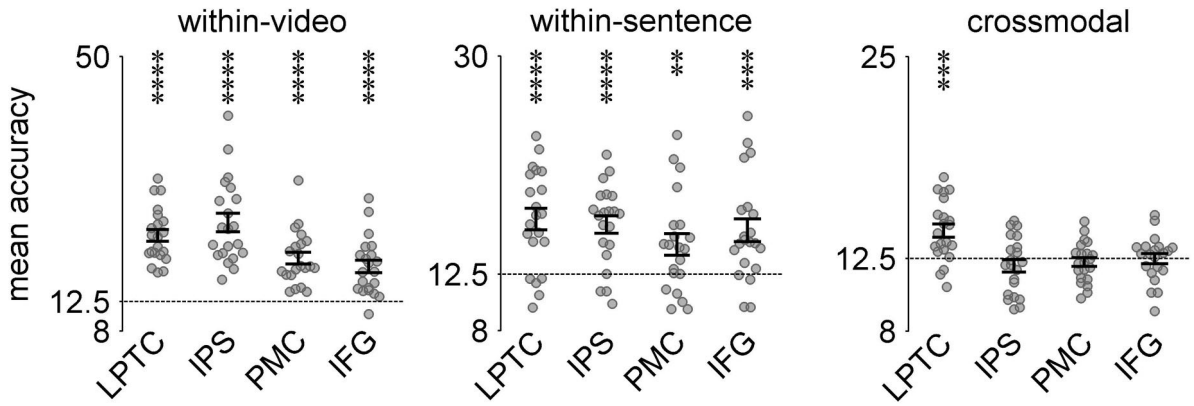
Behavioral results. During fMRI, participants performed a catch trial detection task. In both the video and sentence session, they detected incorrect actions with good accuracy (sentences: 83% ± 2.2 SEM, videos: 80% ± 2.9 SEM). The rate of false alarms was low for all 8 actions (sentences: 0.9% ± 0.3 SEM, videos: 1.5% ± 0.6 SEM) and uncorrelated between the two sessions ($r(6) = -0.32$, $p = 0.44$) suggesting that there were no action-specific similarities in task difficulty/confusability across sessions. A similar result was obtained in the behavioral control experiment, in which participants performed a 2-alternative forced choice task (Supplementary Table 1).

The Post-fMRI ratings for verbalization and visual action imagery revealed, irrespective of session order, no strong tendencies to verbalize during the video session (mean ratings; sentence first: 3.9 ± 0.5 SEM, video first: 2.7 ± 0.6 SEM; on a Likert scale from 1 to 6) and to imagine the actions during the sentence session (sentence first: 3.1 ± 0.4 SEM, video first: 3.8 ± 0.5 SEM; mixed ANOVA interaction: $F(1,19) = 3.71$, $p = 0.07$). However, the correspondence between verbalized actions and sentences during watching the action videos was stronger when the experiment started with the sentence session as compared to starting with the video session. Likewise, the correspondence between imagined actions and actions shown in the videos was higher when the experiment started with the video session as compared to starting with the sentence session (mixed ANOVA interaction: $F(1,19) = 11.31$, $p = 0.003$). The responses to verbalization/imagery ratings and correspondence ratings correlated with each other, i.e., high verbalization ratings were accompanied by high verbalization-sentence correspondence ratings ($r(20) = 0.45$, $p = 0.035$), and high imagery ratings were accompanied by high imagery-video correspondence ratings ($r(20) = 0.57$, $p = 0.009$), Together, these results suggests only weak tendencies to verbalize and to imagine the action across session; but if participants verbalized or imagined the actions, then the verbalized or imagined actions corresponded more strongly to the stimuli they recalled from the first session.

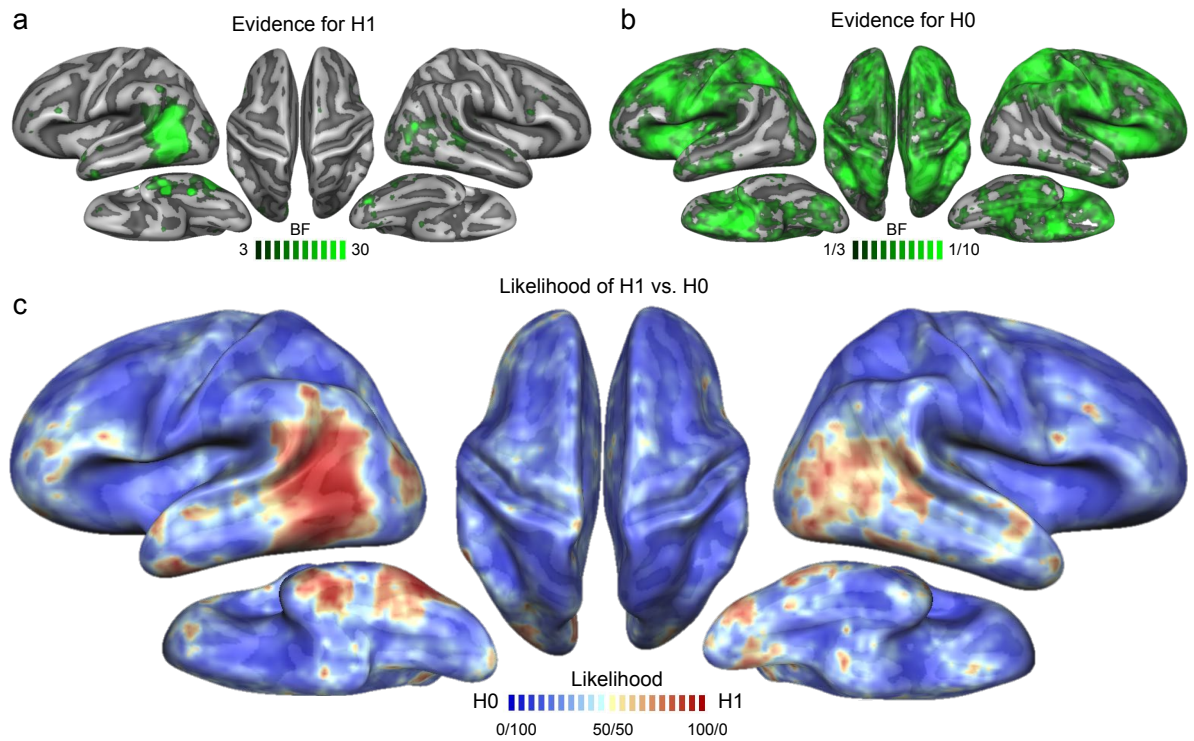
ROI analysis. To investigate differential effects of crossmodal and within-session decoding in frontoparietal and posterior temporal areas in more detail, we performed a ROI analysis. ROIs were selected based on the conjunction (Nichols et al., 2005) of the FDR-corrected RFX contrasts *action videos vs. baseline* and *action sentences vs. baseline*. This conjunction revealed clusters in left inferior frontal gyrus (IFG), left premotor cortex (PMC), left intraparietal sulcus (IPS), and bilateral occipitotemporal cortex, which extended into lateral posterior temporal cortex (LPTC) in the left hemisphere. ROIs were created based on spheres with 12 mm radius around the conjunction peaks within each area (Talairach coordinates x/y/z; IFG: -42/8/22, PMC: -39/-7/37, IPS: -24/-58/40, LPTC: -45/-43/10). From each ROI, classification scheme (crossmodal, within-video, within-sentence), and participant, decoding accuracies were extracted from the searchlight maps and averaged across voxels. Mean decoding accuracies were entered into one-tailed one-sample t-tests and Bayesian

comparisons (Rouder et al., 2009). This analysis revealed FDR-corrected significant effects in all ROIs for the within-video and within-sentence decoding as well as in LPTC for the crossmodal decoding (Supplementary Figure 1, Supplementary Table 3). Bayes factors suggest extreme evidence for above chance decoding (H1; BF > 100) in all ROIs for the within-video decoding, in LPTC and IPS for the within-sentence decoding, and in LPTC for the crossmodal decoding; very strong evidence for H1 (BF = 30-100) in IFG for the within-sentence decoding; and moderate evidence for H1 (BF = 3-10) in PMC for the within-sentence decoding. Moderate evidence for decoding accuracy at chance (H0; BF = 0.33-0.1) was found for the crossmodal decoding in all ROIs expect LPTC (Supplementary Table 3). A repeated measures ANOVA with the factors ROI and DECODING SCHEME revealed a significant interaction ($F(6,120) = 9.99$, $p < 0.0001$) as well as main effects for ROI ($F(3,60) = 18.5$, $p < 0.0001$) and DECODING SCHEME ($F(2,40) = 59.4$, $p < 0.0001$). Two-tailed paired t-tests between LPTC and the other ROIs were all significant (all $t > 3.5$, all $p < 0.002$).

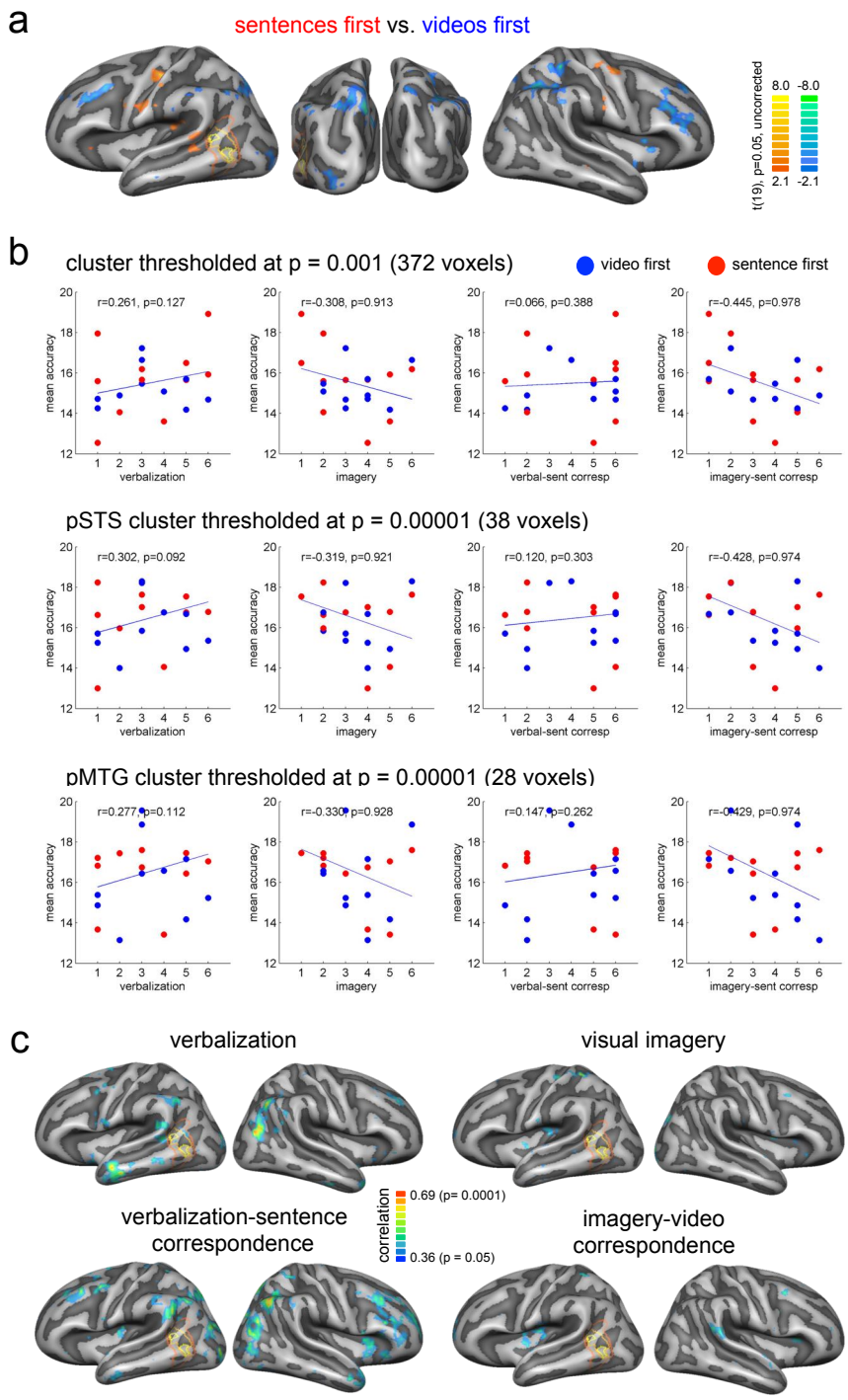
Testing for verbalization and visual imagery effects. Balancing the order of sessions across participants allowed us to investigate whether the crossmodal action decoding in left LPTC might be due to verbalization or visual imagery of the actions. We reasoned that the tendency to verbalize the actions during the video session should be stronger in the participant group beginning with the sentences. By contrast, the tendency to visually imagine the actions should be stronger in the participant group beginning with the videos. Contrasting the decoding maps of the two groups revealed no significant differences in the cluster ($t(20) = 0.04$, $p = 0.97$, two-tailed; Supplementary Fig. 4A). In addition, we found no significant correlations between decoding accuracies and verbalization scores ($r(19) = 0.26$, $p = 0.13$, one-tailed) and imagery scores ($r(19) = -0.31$, $p = 0.97$, one-tailed). As the ROI was relatively large (372 voxels) it could be that a true effect emerging from a subset of voxels in the cluster was averaged out. However, reducing the ROI size by including voxels remaining significant at $p < 0.0001$ (38 voxels in pMTG, 28 voxels in pSTS) revealed no significant effects either (Supplementary Fig. 4B). Correlation analyses using rating scores of the correspondence between verbalized actions and action sentences and between imagined actions and action videos, respectively, revealed no significant effects either (Supplementary Fig. 4B and C). In line with the ROI analyses, a whole brain analysis revealed no substantial overlap between the crossmodal decoding cluster and regions showing correlations between rating scores and decoding accuracies; Supplementary Fig. 2C). Crossmodal action decoding in left LPTC was therefore unlikely to be substantially affected by verbalization during the video session or by visual imagery during the sentence session.



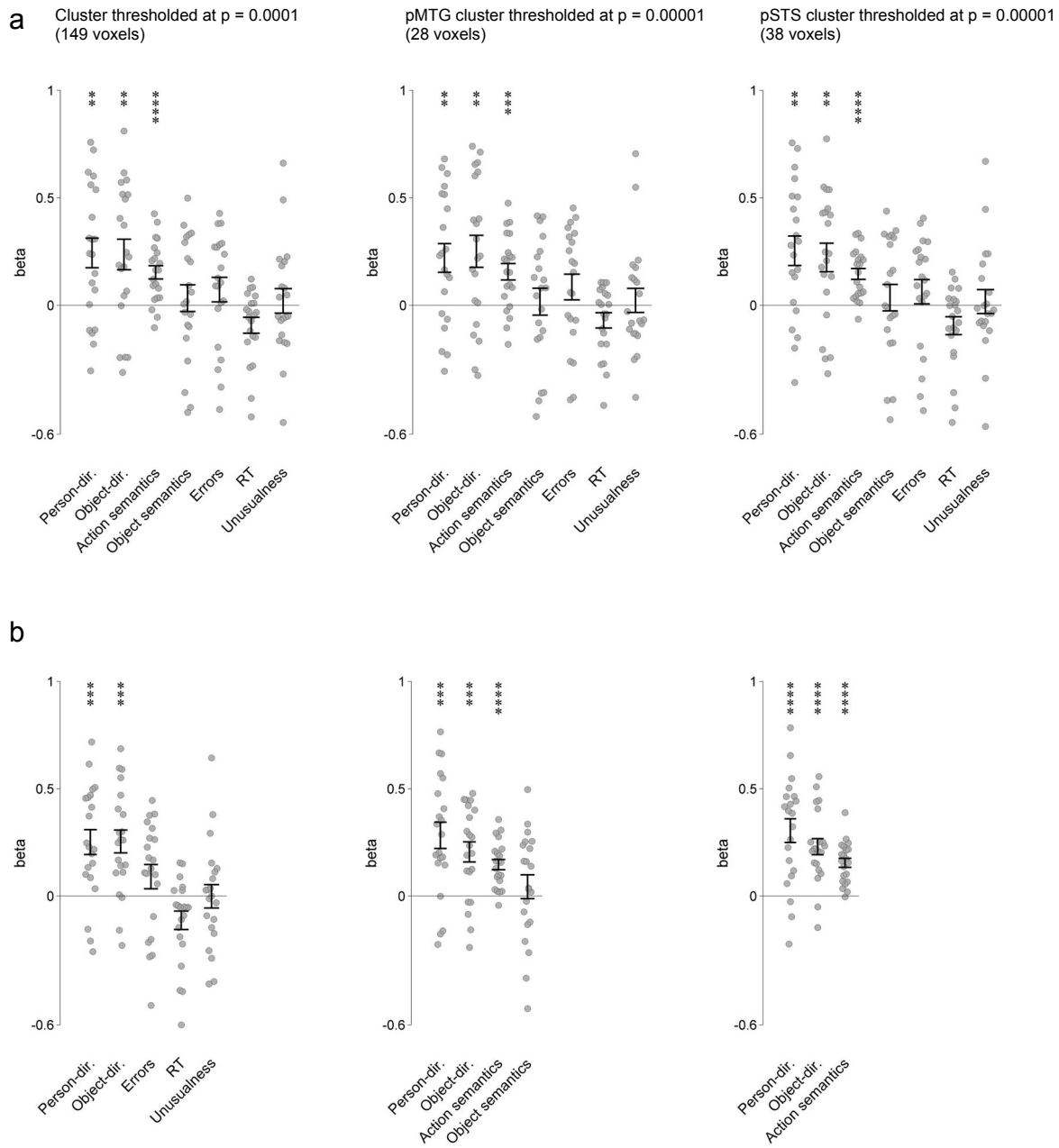
Supplementary Figure 1. Classification accuracies for the within-video, within-sentence, and crossmodal decoding. Asterisks indicate FDR-corrected effects: * $p < 0.05$, ** $p < 0.01$ *** $p = 0.001$, **** $p < 0.0001$. Error bars/lines indicate SEM.



Supplementary Figure 2. Bayesian model comparison for crossmodal action decoding vs. chance. (a) Bayes factors indicating evidence for H1, i.e., the hypothesis that decoding accuracies are above chance. (b) Inverse Bayes factors indicating evidence for H0, i.e., the null hypothesis that decoding accuracies are not above chance. Maps are thresholded at BF = 3 and 1/3, suggesting moderate evidence for H1 and H0, respectively (Jeffreys, 1998). Different upper ends of scales for H1 (30 = very strong evidence) and H0 (10 = strong evidence) maps were chosen to account for asymmetries in ease to find evidence for H1 and H0, respectively. (c) Likelihood map for H1 vs. H0 using a fixed scale (BF/(BF+1)*100).



Supplementary Figure 3. Modulation of crossmodal decoding accuracies by verbalization and visual imagery. (a) Independent two-tailed t test between decoding accuracies of the participant group with session order sentences-videos (sentence first) and the participant group with session order videos-sentences (video first). To reveal any trends of session order effects, maps were leniently thresholded and uncorrected. (b) correlations between decoding accuracies and rating scores for verbalization, visual imagery, correspondence between verbalized actions and sentences, and correspondence between imagined actions and videos (see Methods for details). Decoding accuracies were averaged across voxels that showed significant crossmodal decoding at different thresholds. Blue dots indicate participants of the "video first" group, red dots indicate participants of the "sentence first" group. (c) Whole brain maps of correlations between rating scores and crossmodal decoding accuracies.



Supplementary Figure 4. Robustness of crossmodal multiple regression RSA effects across different ROI sizes (a) and models used in the multiple regression RSA (b). Asterisks indicate FDR-corrected effects: * $p < 0.05$, ** $p < 0.01$ *** $p = 0.001$, **** $p < 0.0001$. Error bars/lines indicate SEM.

Supplementary Table 1. Clusters identified in the crossmodal, within-sentence, and within-video action decoding.

Region	x	y	z	t	p	Accuracy
crossmodal						
left pMTG	-54	-61	4	7.78	8.94E-08	17.7
left pSTS	-48	-49	10	6.91	1.79E-07	16.9
within video						
left LOTC	-45	-64	1	18.24	3.11E-14	35.5
right LOTC	45	-58	4	15.92	3.99E-13	35.0
left aIPL	-54	-28	31	14.61	1.95E-12	31.0
right aIPL	54	-22	31	13.11	1.40E-11	27.5
right IPS	24	-55	43	12.70	2.50E-11	31.1
left IPS	-27	-79	22	11.00	3.09E-10	30.4
within sentences						
left LOTC	-45	-55	1	7.03	1.00E-06	21.1
right LOTC	45	-55	-2	7.59	1.31E-07	18.3
left IPS	-27	-58	43	7.00	1.00E-06	19.7
right IPS	12	-70	34	8.09	4.93E-08	19.1
left LOC	-21	-85	1	11.30	1.96E-10	26.9
right LOC	24	-82	1	9.36	4.75E-09	23.9
left pSTG	-45	-40	31	5.50	2.20E-05	20.5
left VWFA	-42	-52	-20	6.74	1.00E-06	19.7
left IFG/IFS	-42	23	22	5.48	2.30E-05	19.5
right IFG/MFG	-42	23	22	5.48	2.30E-05	19.5
left PMC	-54	2	31	5.34	3.10E-05	19.3
right IFJ	33	8	34	6.96	1.00E-06	18.0
left ATL	-60	-13	-5	4.81	1.06E-04	18.3
right ATL	60	-13	-2	4.00	7.12E-04	17.8

Coordinates (x, y, z) in Talairach space. Decoding accuracy at chance is 12.5%. Maps were TFCE corrected. For the within video and within sentence action decoding, only main clusters with distinctive peaks are reported. Abbreviations: aIPL, anterior inferior parietal lobe; ATL, anterior temporal lobe; IFG, inferior frontal gyrus; IFJ, inferior frontal junction; IFS, inferior frontal sulcus; IPS, intraparietal sulcus; LOC, lateral occipital cortex; LOTC, lateral occipitotemporal cortex; MFG, middle frontal gyrus; PMC, premotor cortex; pMTG, posterior middle temporal gyrus; pSTG, posterior superior temporal gyrus ; pSTS, posterior superior temporal sulcus; VWFA, visual word form area.

Supplementary Table 2. Behavioral results of the fMRI experiment, the behavioral (two-alternatives forced choice) control experiment, and the rating (using Likert scales from 1 = not at all to 6 = very much; see Methods for details).

	open	close	give	take	stroke	scratch	agree	disagree	catch trials
fMRI sentence session									
mean hit/CR rate	0.990	0.992	0.973	0.999	0.992	0.994	0.995	0.996	0.832
SEM	0.003	0.003	0.005	0.001	0.003	0.003	0.003	0.003	0.022
fMRI video session									
mean hit/CR rate	0.988	0.981	0.998	0.991	0.959	0.987	0.993	0.978	0.805
SEM	0.003	0.008	0.001	0.004	0.016	0.006	0.004	0.012	0.029
2AFC sentence session									
mean accuracy	0.986	0.993	0.972	0.979	0.979	0.993	1.000	1.000	0.819
SEM	0.009	0.007	0.021	0.011	0.015	0.007	0.000	0.000	0.056
mean RT	1961	1968	2044	1963	1973	1954	1986	1972	2121
SEM	91	98	106	100	83	101	77	78	73
2AFC video session									
mean accuracy	0.986	1.000	1.000	1.000	0.924	0.972	0.993	1.000	0.875
SEM	0.009	0.000	0.000	0.000	0.024	0.012	0.007	0.000	0.035
mean RT	1923	1903	1819	1821	1753	1689	1670	1657	1897
SEM	66	69	71	67	93	90	85	80	70
Unusualness sentences									
mean rating	1.09	1.36	1.18	1.09	2.09	1.18	1.27	1.45	NA
SEM	0.09	0.28	0.12	0.09	0.37	0.12	0.19	0.25	NA
Unusualness videos									
mean rating	1.18	1.00	1.36	1.36	2.18	1.45	1.91	2.00	NA
SEM	0.12	0.00	0.20	0.28	0.44	0.28	0.31	0.36	NA
Person-directedness									
mean rating	1.15	1.25	4.20	3.40	1.20	1.15	5.50	5.65	NA
SEM	0.08	0.12	0.39	0.41	0.12	0.11	0.24	0.17	NA
Object-directedness									
mean rating	5.80	5.80	5.65	5.75	1.65	1.70	1.15	1.15	NA
SEM	0.09	0.12	0.17	0.12	0.32	0.35	0.11	0.11	NA

CR, correct rejections; 2AFC, Two-alternative forced choice.

Supplementary Table 3. One tailed t tests and Bayesian comparisons for within-video, within-sentence, and crossmodal decoding.

	within-video			within-sentence			crossmodal		
	t(20)	p	BF	t(20)	p	BF	t(20)	p	BF
LPTC	11.18	<0.0001	>1000	5.10	<0.0001	953.27	4.17	0.0002	138.61
IPS	8.56	<0.0001	>1000	5.64	<0.0001	>1000	-1.20	0.8773	0.11
PMC	7.33	<0.0001	>1000	2.78	0.0058	8.88	-0.85	0.7984	0.13
IFG	5.55	<0.0001	>1000	3.91	0.0004	81.31	-0.08	0.5314	0.21