An architecture for encoding sentence meaning in left mid-superior temporal cortex

Steven M. Frankland[a,b,1] and Joshua D. Greene[a,b]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138; and [b]Center for Brain Science, Harvard University, Cambridge, MA 02138

Edited by Stanislas Dehaene, INSERM U992CEA/Saclay, College de France, Gif/Yvette, France, and approved July 24, 2015 (received for review December 2, 2014)

Human brains flexibly combine the meanings of words to compose structured thoughts. For example, by combining the meanings of "bite," "dog," and "man," we can think about a dog biting a man, or a man biting a dog. Here, in two functional magnetic resonance imaging (fMRI) experiments using multivoxel pattern analysis (MVPA), we identify a region of left mid-superior temporal cortex (lmSTC) that flexibly encodes "who did what to whom" in visually presented sentences. We find that lmSTC represents the current values of abstract semantic variables ("Who did it?" and "To whom was it done?") in distinct subregions. Experiment 1 first identifies a broad region of lmSTC whose activity patterns (*i*) facilitate decoding of structure-dependent sentence meaning ("Who did what to whom?") and (*ii*) predict affect-related amygdala responses that depend on this information (e.g., "the baby kicked the grandfather" vs. "the grandfather kicked the baby"). Experiment 2 then identifies distinct, but neighboring, subregions of lmSTC whose activity patterns carry information about the identity of the current "agent" ("Who did it?") and the current "patient" ("To whom was it done?"). These neighboring subregions lie along the upper bank of the superior temporal sulcus and the lateral bank of the superior temporal gyrus, respectively. At a high level, these regions may function like topographically defined data registers, encoding the fluctuating values of abstract semantic variables. This functional architecture, which in key respects resembles that of a classical computer, may play a critical role in enabling humans to flexibly generate complex thoughts.

fMRI | cognitive architecture | compositionality | comprehension | PBEC

Yesterday, the world's tallest woman was serenaded by 30 pink elephants. The previous sentence is false, but perfectly comprehensible, despite the improbability of the situation it describes. It is comprehensible because the human mind can flexibly combine the meanings of individual words ("woman," "serenade," "elephants," etc.) to compose structured thoughts, such as the meaning of the aforementioned sentence (1, 2). How the brain accomplishes this remarkable feat remains a central, but unanswered, question in cognitive science.

Given the vast number of sentences we can understand and produce, it would be implausible for the brain to allocate individual neurons to represent each possible sentence meaning. Instead, it is likely that the brain employs a system for flexibly combining representations of simpler meanings to compose more complex meanings. By "flexibly," we mean that the same meanings can be combined in many different ways to produce many distinct complex meanings. How the brain flexibly composes complex, structured meanings out of simpler ones is a matter of long-standing debate (3–10).

At the cognitive level, theorists have held that the mind encodes sentence-level meaning by explicitly representing and updating the values of abstract semantic variables (3, 5) in a manner analogous to that of a classical computer. Such semantic variables correspond to basic, recurring questions of meaning such as "Who did it?" and "To whom was it done?" On such a view, the meaning of a simple sentence is partly represented by filling in these variables with representations of the appropriate semantic components. For example, "the dog bit the man" would be built out of the same semantic components as "the man bit the dog," but with a reversal in the values of the "agent" variable ("Who did it?") and the "patient" variable ("To whom was it done?"). Whether and how the human brain does this remains unknown.

Previous research has implicated a network of cortical regions in high-level semantic processing. Many of these regions surround the left sylvian fissure (11–19), including regions of the inferior frontal cortex (13, 14), inferior parietal lobe (12, 20), much of the superior temporal sulcus and gyrus (12, 15, 21), and the anterior temporal lobes (17, 20, 22). Here, we describe two functional magnetic resonance imaging (fMRI) experiments aimed at understanding how the brain (in these regions or elsewhere) flexibly encodes the meanings of sentences involving an agent ("Who did it?"), an action ("What was done?"), and a patient ("To whom was it done?").

First, experiment 1 aims to identify regions that encode structure-dependent meaning. Here, we search for regions that differentiate between pairs of visually presented sentences, where these sentences convey different meanings using the same words (as in "man bites dog" and "dog bites man"). Experiment 1 identifies a region of left mid-superior temporal cortex (lmSTC) encoding structure-dependent meaning. Experiment 2 then asks how the lmSTC represents structure-dependent meaning. Specifically, we test the long-standing hypothesis that the brain represents and updates the values of abstract semantic variables (3, 5): here, the agent ("Who did it?") and the patient ("To whom was it done?"). We search for distinct neural populations in lmSTC that encode these variables, analogous to the data registers of a computer (5).

Experiment 1

In experiment 1, subjects undergoing fMRI read sentences describing simple events. Each sentence expressed a meaning, or

Significance

The 18th-century Prussian philosopher Wilhelm von Humbolt famously noted that natural language makes "infinite use of finite means." By this, he meant that language deploys a finite set of words to express an effectively infinite set of ideas. As the seat of both language and thought, the human brain must be capable of rapidly encoding the multitude of thoughts that a sentence could convey. How does this work? Here, we find evidence supporting a long-standing conjecture of cognitive science: that the human brain encodes the meanings of simple sentences much like a computer, with distinct neural populations representing answers to basic questions of meaning such as "Who did it?" and "To whom was it done?"

Author contributions: S.M.F. and J.D.G. designed research; S.M.F. performed research; S.M.F. and J.D.G. contributed new reagents/analytic tools; S.M.F. analyzed data; and S.M.F. and J.D.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: franklan@fas.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1421236112/-/DCSupplemental.

"proposition," which could be conveyed in either the active or passive voice (e.g., "the ball hit the truck"/"the truck was hit by the ball"). Each such sentence could be reversed to yield a mirror image proposition (e.g., "the truck hit the ball"/"the ball was hit by the truck"), which was also included in the stimulus set. We call these "mirror image proposition pairs." Members of these pairs contain the same words and have the same syntactic structure, but the words are differentially assigned to the agent and patient roles to form different sentence-level meanings.

A region encoding the meanings of these sentences should have the following two properties. First, patterns of activity in such a region should differentially encode members of mirror image propositions pairs. For example, the propositions conveyed by "the truck hit the ball" and "the ball hit the truck" should elicit distinct patterns of activity. Second, the instantiation of such patterns should predict downstream neural responses that depend on understanding "who did what to whom." For example, patterns related to sentence-level meaning should predict differential affective responses to "the grandfather kicked the baby" and "the baby kicked the grandfather." Experiment 1 used two key analyses, corresponding to these two functional properties. First, we applied multivoxel pattern analysis (23–25) and a whole-brain searchlight procedure (26) to identify sets of contiguous voxels that distinguish between members of mirror image proposition pairs. Second, we developed a pattern-based effective connectivity (PBEC) analysis to determine whether patterns related to affectively salient sentences (e.g., "the grandfather kicked the baby") mediate the relationship between the sentence presented and affective responses elsewhere in the brain. Jointly, these analyses establish candidate regions for encoding structure-dependent meaning that can be further probed in experiment 2.

Whole-Brain Searchlight Analysis. First, using a linear classifier, we searched for regions whose patterns of activity distinguished between members of mirror image proposition pairs: for example, between the proposition conveyed by "the truck hit the ball" (as well as "the ball was hit by the truck") and the proposition conveyed by "the ball hit the truck" (as well as "the truck was hit by the ball"). The use of mirror image propositions ensures that basic lexico-semantic content, syntactic structure, and summed word frequency are matched between the propositions to be discriminated. Active and passive forms of each proposition were treated as identical in all analyses, allowing us to identify underlying semantic representations, controlling for visual features of the stimuli and surface syntax. All propositions were presented separately, and multiple times, to better estimate the pattern of activity evoked by each proposition. For experiment 1, classifiers were thus tested on their ability to discriminate between new tokens of the mirror image propositions on which they were trained.

For this initial searchlight analysis, we used four mirror image pairs of propositions, two involving animate entities and two involving inanimate entities. For each subject ($n = 16$), we averaged classification accuracies across these four pairwise classification problems to yield a map of the mean classification accuracy by region. Group-level analysis identified a region of lmSTC ($k = 123$; Talairach center: −59, −25, 6) that reliably distinguished between mirror image propositions ($P < 0.0001$, corrected; mean accuracy, 57%) (see left temporal region in Fig. 1). This result was not driven by a particular subset of the stimuli (*Supporting Information*). A second significant cluster was discovered along the right posterior insula/extreme capsule region ($P < 0.001$, corrected; 37, −9, 6; mean accuracy, 56.4%). However, this second region failed to meet additional, minimal functional criteria for encoding sentence meaning (*Supporting Information*).

PBEC Analysis. The foregoing searchlight analysis suggests that lmSTC represents critical aspects of sentence-level meaning. If this hypothesis is correct, then the particular pattern instantiated in lmSTC should also predict downstream neural responses when those responses depend on an understanding of "who did what to whom." Our second analysis in experiment 1 attempts to determine whether the patterns of activity in lmSTC predict affective neural responses elsewhere in the brain.

To test this hypothesis, we used, within the same experiment, an independent set of mirror image proposition pairs in which one proposition is more affectively salient than its counterpart, as in "the grandfather kicked the baby" and "the baby kicked the grandfather." (Differences in affective salience were verified with independent behavioral testing. See *Supporting Information*.) We predicted that patterns of activity in lmSTC (as delineated by the independent searchlight analysis) would statistically mediate the relationship between the sentence presented and the affective neural response, consistent with a causal relationship (27). This PBEC analysis proceeded in three steps.

First, we confirmed that patterns of activity in the region of lmSTC identified by the searchlight analysis can discriminate between these new mirror image propositions [$t_{(15)} = 3.2$; $P = 0.005$; mean accuracy, 58.3%], thus replicating the above findings with new stimuli. Second, we identified brain regions that respond more strongly to affectively salient propositions (e.g., "the grandfather kicked the baby" > "the baby kicked the grandfather"). This univariate contrast yielded effects in two brain regions, the left amygdala (−28, −7, −18) and superior parietal lobe (−38, −67, 47), ($P < 0.001$, corrected). Given its well-known role in affective processing (28), we interpreted this amygdala response as an affective signal and focused on this region in our subsequent mediation analysis. Third, and most critically, we examined the relationship between patterns of activity in lmSTC and the magnitude of the amygdala's response. The first of the above analyses shows that "the grandfather kicked the baby" produces a different pattern in lmSTC than "the baby kicked the grandfather" (etc.). If these patterns actually reflect structure-dependent meaning, then these patterns should mediate the relationship between the sentence presented and the amygdala's response on a trial-by-trial basis.

To quantify the pattern of activity in lmSTC on each trial, we used the signed distance of each test pattern from the classifier's decision boundary (*Supporting Information*). This signed distance variable reflects the content of the classifier's decision regarding the sentence (the sign), as well as what one may think of as its "confidence" in that decision (the distance). According to our hypothesis, trials in which the pattern is confidently classified as "the grandfather kicked the baby" (etc.), rather than "the baby kicked the grandfather" (etc.), should be trials in which the amygdala's response is robust. Here, we are supposing that the classifier's "confidence" will reflect the robustness of the semantic representation, which in turn may influence downstream affective responses in the amygdala.

As predicted, the pattern of activity instantiated in lmSTC predicted the amygdala's response [$t_{(15)} = 3.96$, $P = 0.0013$], over and above both the mean signal in lmSTC and the content of the stimulus. The pattern of activity in the lmSTC explains unique variance in the amygdala's response, consistent with a causal model whereby information flows from the sentence on the screen, to a pattern of activity in the lmSTC, to the amygdala [$P < 0.01$, by Monte Carlo simulation (29, 30); Sobel test (27), $z = 2.47$, $P = 0.013$] (Fig. 1). The alternative model reversing the direction of causation between the lmSTC and amygdala was not significant (Monte Carlo, $P > 0.10$; Sobel, $z = 1.43$, $P = 0.15$), further supporting the proposed model.

There are several possible sources of trial-to-trial variability in lmSTC's responses (see *Supporting Information* for more discussion). For example, a participant's inattention might disrupt the semantic representation in lmSTC, making the trial more difficult to classify and, at the same time, making the amygdala

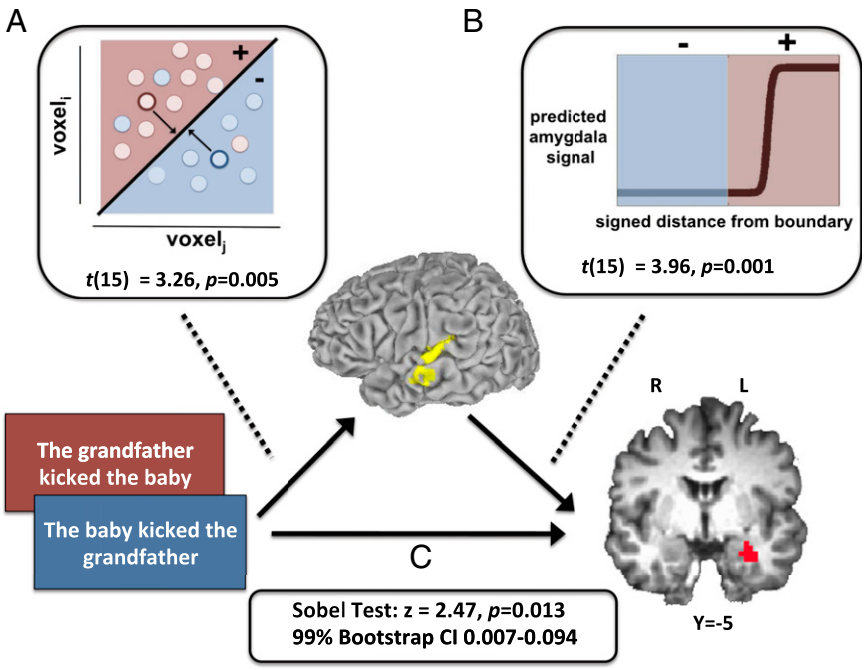

Fig. 1. Model of information flow from stimulus to lmSTC to amygdala in experiment 1. (*A*) A pattern classifier determines which of two propositions was presented using activity in lmSTC. Distance from the classification boundary indicates the extent to which a learned pattern was instantiated. The red region corresponds to the emotionally evocative proposition (e.g., "the grandfather kicked baby"), whereas blue corresponds to the less evocative proposition ("the baby kicked grandfather"). (*B*) For each trial, the classifier's signed distance from the classification boundary was transformed by a sigmoidal function and used to predict the mean level of activity in the left amygdala. (*C*) Patterns in lmSTC mediate the relationship between the proposition on the screen and the amygdala's response, consistent with a model according to which the lmSTC encodes the structured representations necessary to generate an emotional response.

response weaker than otherwise expected. Regardless of the source of the variation in these patterns, the present data provide evidence that neural representations of structure-dependent meanings in lmSTC predict downstream affective responses, consistent with our causal model.

Thus, experiment 1 shows that a region of lmSTC meets our two initial functional criteria for a region encoding structure-dependent sentence meaning. First, its patterns of activity differentiate between mirror image propositions containing the same words and syntactic structure. Second, these patterns statistically mediate the relationship between the sentence presented and affective neural responses that depend on understanding "who did what to whom." Experiment 1 does not, however, explain how this region encodes such information. Experiment 2 aims to further validate the results of experiment 1 and to illuminate the mechanism by which this region encodes these structure-dependent meanings.

Experiment 2

In experiment 2, we test the hypothesis that lmSTC flexibly encodes these meanings (at least in part) by explicitly representing the values of the agent ("Who did it?") and the patient ("To whom was it done?") (5). To evaluate this possibility, we searched for subregions of lmSTC whose patterns of activity reflect the current value of these variables. We performed separate searches for each variable, searching for subregions encoding "Who did it?" and "To whom it was done?" across verb contexts. Thus, we aimed to identify regions that are specialized for representing the agent and patient variables as such.

Experiment 2 ($n = 25$) used a stimulus set in which four nouns ("man," "girl," "dog," and "cat") were assigned to the agent and patient roles for each of five verbs ("chased," "scratched," etc.), in both active and passive forms (Fig. 2*A*). Thus, subjects undergoing fMRI read sentences such as "the dog chased the man" and "the girl was scratched by the cat," exhausting all meaningful combinations, excluding combinations assigning the same noun to both roles (e.g., "the man chased the man").

We acquired partial-volume, high-resolution (1.5-mm^3 isotropic voxels) functional images covering the lmSTC. We used separate searchlight analyses within each subject to identify subregions of lmSTC encoding information about the identity of the agent or patient (Fig. 2 *B* and *C*). For our principal searchlight analyses, four-way classifiers were trained to identify the agent or patient using data generated by four out of five verbs. The classifiers were then tested on data from sentences containing the withheld verb. For example, the classifiers were tested using patterns generated by "the dog chased the man," having never previously encountered patterns generated by sentences involving "chased," but having been trained to identify "dog" as the agent and "man" as the patient in other verb contexts. This procedure was repeated holding each verb's data out of the training set, and the results were averaged across cross-validation iterations. Thus, this analysis targets regions that instantiate consistent patterns of activity for (for example) "dog as agent" across verb contexts, discriminable from "man as agent" (and likewise for other nouns). A region that carries this information therefore encodes "Who did it?" across the nouns and verb contexts tested. This same procedure was repeated to decode the identity of the patient.

These searchlight analyses revealed distinct subregions of lmSTC that reliably carry information about the identity of the agent and the patient (Fig. 3*A*). Within the anterior portion of lmSTC, a medial subregion located on the upper bank of the superior temporal sulcus (STS) encoded information about the identity of the agent ($P < 0.01$, corrected; −46, −18, 1). A spatially distinct lateral subregion, encompassing part of the upper bank of the STS, as well as the lateral superior temporal gyrus (STG) carried patient information ($P < 0.0001$, corrected; −57, −10, 2) across subjects. These anterior agent and patient clusters are adjacent, but nonoverlapping in this analysis. A follow-up analysis using independent data to define each participant's agent and patient clusters found that these subregions are significantly dissociable by their informational content [$F_{\text{Region}\times\text{Content}(1,24)} = 12.99$, $P_{\text{perm}} = 0.001$] (Fig. 3). This searchlight analysis also revealed a second agent cluster, posterior and superior to the clusters described above, located primarily within the posterior STS ($P < 0.02$, corrected; −57, −37, 7). Post hoc analyses found the classification accuracies driving these results to be only modestly above chance levels of 25%, but statistically reliable across our set of 25 subjects (mean accuracies across subjects: anterior agent, 27.1%; posterior agent, 28.1%; patient, 26.6%).

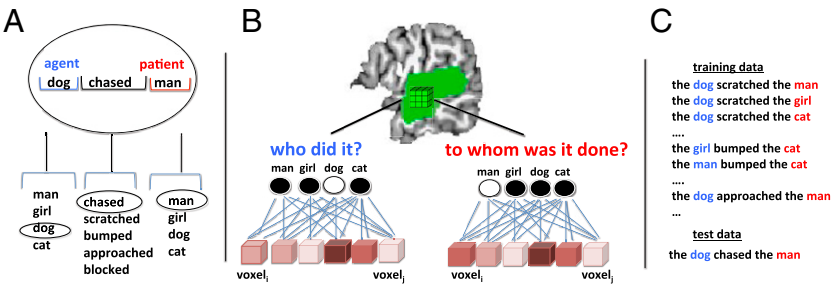

Fig. 2. Experiment 2 design. (*A*) Subjects read sentences constructed from a menu of five verbs and four nouns, with one noun in the agent role and another in the patient role. (*B*) For each trial, separate pattern classifiers attempted to identify the agent and the patient based on activity within subregions of lmSTC. (*C*) Classifiers were trained using data from four of five verbs and tested on data from the withheld verb. This required the classifiers to identify agents and patients based on patterns that are reused across contexts.

As in experiment 1, post hoc analyses ruled out the possibility that these results were driven by a subset of items, as these regions were relatively consistent in their ability to discriminate between particular pairs of nouns and to generalize across the five verb contexts. (See *Supporting Information* for detailed procedures and results for these post hoc analyses.) These results thus suggest that the regions identified by the experiment 2 searchlight analyses are generally involved in encoding noun–role bindings across the nouns and verbs used. No regions of lmSTC carried information about the surface subject and surface object of the sentence. For example, no lmSTC region encoded "the dog chased the man" and "the dog was chased by the man" as similar to each other, but different from "the man chased the dog" and "the man was chased by the dog." Within lmSTC, the encoding appears, instead, to be based on deeper semantics, encoding the underlying agent and patient of the sentence, independent of which noun serves as the sentence's surface subject or object, consistent with experiment 1.

These findings provide preliminary evidence that these subregions of lmSTC encode the values of the agent and patient variables. However, it remains open whether and to what extent these subregions are specialized for representing agent and patient information—that is, whether they tend to represent one kind of information and not the other. To address this question, we conducted planned post hoc analyses that separately defined agent and patient regions within each subject using data from the remaining subjects. We assessed the significance of these effects using both conventional parametric statistics and permutation tests (*Supporting Information*). Within subjects' independently localized patient regions, patient identification accuracy was significantly greater than agent identification accuracy across subjects [lateral lmSTC: $t_{(24)} = 2.96$, $P = 0.006$; permutation test: 0.006]. Within the posterior agent region, agent identification was significantly above chance [$t_{(24)} = 2.38$, $P = 0.01$; permutation test: $P = 0.008$]. Within the anterior agent region, the classification effect was somewhat weaker [$t_{(24)} = 2.04$, $P = 0.02$; permutation test: $P = 0.055$]. As expected, patient identification was at chance in both the anterior agent region [$t_{(24)} = 0.86$, $P = 0.2$; permutation test: $P = 0.22$] and the posterior agent region [$t_{(24)} = -0.29$, $P = 0.39$; permutation test: $P = 0.38$]. However, the direct comparison of accuracy levels for agent and patient identification was not statistically significant in the anterior agent region ($P = 0.27$; permutation test: $P = 0.26$) or the posterior agent region ($P = 0.15$; permutation test: $P = 0.15$). See Fig. 3*B*.

To further assess the role specificity of these subregions, we localized a large portion of the anterior lmSTC in a manner that was unbiased with respect to its role preference, and then quantified the average preferences of slices of voxels at each X coordinate (*Supporting Information*). We found a clear trend in role preference along the medial-lateral axis, with medial portions preferentially encoding agent information and lateral portions preferentially encoding patient information (Fig. 3*C*). From the present data, we cannot determine whether the observed graded shift in role preference exists within individuals, or

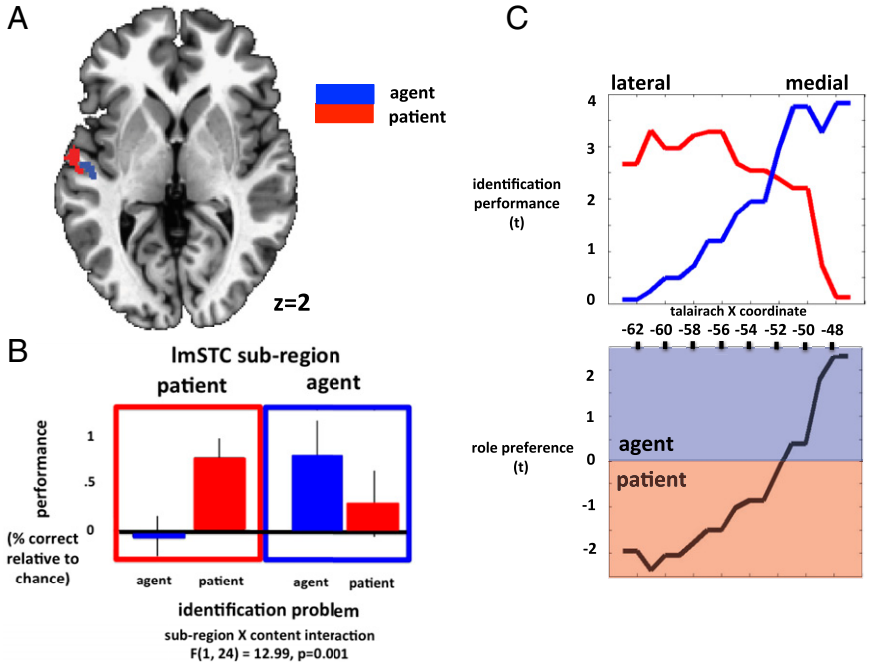

Fig. 3. (*A*) Searchlight analyses identified adjacent, but nonoverlapping subregions of anterior lmSTC that reliably encoded information about agent identity (medial, blue) and patient identity (lateral, red). (*B*) Post hoc analyses find that these adjacent regions differ significantly in the information they encode. These analyses define each subject's agent and patient subregions using data from other subjects, and the statistics computed within each subject's agent/patient region reflect the average accuracy of all voxel neighborhoods across that region. (*C*) Across subjects, medial portions of anterior lmSTC preferentially encode agent information, whereas lateral portions of anterior lmSTC preferentially encode patient information.

simply results from averaging across individuals exhibiting more abrupt transitions.

A final searchlight analysis within lmSTC identified two additional subregions supporting identification of the present verb (*Supporting Information*). The anterior verb subregion ($P < 0.025$; −61, −15, 2) was adjacent to the patient subregion. The posterior verb subregion ($P < 0.0001$; −55, −49, 5) in the posterior STS partially overlapped with the posterior agent region.

The foregoing analyses strongly suggest that a lateral subregion of anterior lmSTC selectively encodes information about the identity of the current patient, and somewhat less strongly, that a medial portion of anterior lmSTC selectively encodes information about the identity of the current agent. In addition, we identified two subregions of lmSTC supporting classification of the verb present on a given trial (*Supporting Information*). Together, these results indicate that distinct subregions of lmSTC separately and dynamically represent the semantic information sufficient to compose complex representations involving an agent, a patient, and an action.

A third experiment replicates the findings of experiment 2. Once again, we find that a medial region of lmSTC encodes information about the agent while a neighboring lateral region encodes information about the patient (*Supporting Information*).

Discussion

The experiments presented here begin to address an important unanswered question in cognitive neuroscience (2–6): How does the brain flexibly compose structured thoughts out of simpler ideas? We provide preliminary evidence for a long-standing theoretical conjecture of cognitive science: that the brain, on some level, functions like a classical computer, representing structured semantic combinations by explicitly encoding the values of abstract variables (3, 5). Moreover, we find evidence that the agent and patient variables are topographically represented across the upper bank of the left STS and lateral STG, such that adjacent cortical regions are differentially involved in encoding the identity of the agent and patient. At a high level, these regions may be thought of as functioning like the data registers of a computer, in which time-varying activity patterns temporarily represent the current values of these variables (5). This functional architecture could support the compositional encoding of sentence meaning involving an agent and a patient, as these representations can be simultaneously instantiated in adjacent regions to form complex representations with explicit, constituent structure. These structured representations may in turn be read by other neural systems that enable reasoning, decision making, and other high-level cognitive functions.

The present results are broadly consistent with previous research concerning the neural loci of sentence-level semantic processing while, at the same time, offering new insight into how such semantic information is represented. With respect to functional localization, previous research has implicated the lmSTC in phrase and sentence-level semantic processing using both functional neuroimaging and lesion data (11–13, 15, 18, 21). However, lmSTC is by no means the only region consistently implicated in higher-order semantic processing, as research has reliably documented the involvement of the anterior regions of the temporal lobe (20, 22), left inferior parietal lobe (12, 20), and left inferior frontal cortices (13, 14). The two studies presented here suggest that lmSTC may be more narrowly involved in encoding the values of semantic role variables. This narrower claim is consistent with multiple pieces of preexisting experimental evidence.

First, fMRI studies (15, 31) have found increased activation in a similar region of mid-left STG/STS in response to implausible noun–verb combinations that violate a verb's selectional restrictions (e.g., "the thunderstorm was ironed") (but see ref. 32 for conflicting results). More directly, an fMRI study (21) finds that the repetition of a sentence's meaning produces adaptation effects in the lmSTC, even when that meaning is expressed using different surface syntactic forms, such as the active and passive voice. These semantic adaptation effects occur in mid-STG and middorsal MTG/ventral STS when sentences are presented aurally, and in middorsal MTG/midventral STS when presented visually. Finally, and perhaps of most direct relevance, patients with damage to lmSTC have been found to have specific deficits in determining "who did what to whom" in response to both sentences and visual scenes representing actions (11). Here, the locus of damage that most consistently predicts impaired performance across tasks appears to correspond to the anterior subregion of lmSTC in which we find the agent and patient variables to be topographically represented.

The present results build on this literature and extend our understanding in several key ways. First, experiment 1 uses multivariate methods to demonstrate that lmSTC carries information about sentence-level meaning. Second, experiment 1 employs a PBEC analysis to link these patterns of activity to affect-related amygdala responses, consistent with a model whereby lmSTC enables the comprehension necessary to produce an appropriate affective response to a morally salient sentence. Third, and most critically, experiment 2 provides insight into how the lmSTC encodes sentence-level meaning, namely by representing the values of the agent and patient variables in spatially distinct neural populations.

Given that the present results were generated using only linguistic stimuli, the current data are silent as to whether these representations are part of a general, amodal "language of thought" (33), or whether they are specifically linguistic. In particular, it is not known whether results would be similar using alternative modes of presentation, such as pictures. We note that the aforementioned lesion study of ref. 11 reports deficits in comprehension of pictorial stimuli following damage to this region. However, linguistic deficits could disrupt comprehension of pictures if pictorial information is normally translated into words. Although such questions remain open, we emphasize that the representations examined here are related to the underlying semantic properties of our stimuli, for reasons explained in detail above. They encode information that would have to be encoded, in some form, by any semantic system capable of supporting genuine comprehension.

In evaluating the significance of the present results, we note that the classification accuracies observed here are rather modest. Thus, we are by no means claiming that it is now possible to "read" people's thoughts using patterns of activity in lmSTC. Nor are we claiming that the lmSTC is the unique locus of complex thought. On the contrary, we suspect that the lmSTC is merely part of a distributed neural system responsible for accessing and combining representations housed elsewhere in the cortex (10). We regard the observed effects as significant, not because of their size, but because they provide evidence for a distinctive theory of high-level semantic representation. We find evidence for a functional segregation, and corresponding spatial segregation, based on semantic role, which may enable the composition of complex semantic representations. Such functional segregation need not take the form of spatial segregation, but insofar as it does, it becomes possible to provide evidence for functional segregation using fMRI, as done here.

A prominent alternative model for the encoding of complex meanings holds that binding is signaled through the synchronization (or desynchronization) of the firing phases of neurons encoding a complex representation's constituent semantic elements (6–8). Given the limited temporal resolution of fMRI, the current design cannot provide direct evidence for or against temporal synchrony as a binding mechanism. However, the present data suggest that such temporal correlations may be unnecessary in this case, because these bindings may instead be encoded through the instantiation of distributed patterns of activity in spatially dissociable patches of cortex devoted to representing distinct semantic variables. Nevertheless, it is possible

that temporal synchrony plays a role in these processes. Another alternative class of models posits the use of matrix operations to combine spatially distributed representations into conjunctive representations (e.g., "man as agent") (4, 34). Although such models do not necessarily predict the current results, they could potentially be augmented to accommodate them, incorporating separate banks of neurons that encode conjunctive representations for distinct semantic roles. This anatomical strategy, in which separate banks of neurons represent different semantic role variables, is used and expanded in a recent computational model of variable binding that mimics the capacities and limitations of human performance (10). This biologically plausible model employs representations that function like the pointers used in some computer programming languages. It is possible that the patterns of activity within the agent and patient regions that we identify here likewise serve as pointers to richer representations housed elsewhere in cortex.

Although the present work concerns only one type of structured semantic representation (simple agent–verb–patient combinations) and one mode of presentation (visually presented sentences), it supports an intriguing possibility (5): that the explicit representation of abstract semantic variables in distinct neural circuits plays a critical role in enabling human brains to compose complex ideas out of simpler ones.

Materials and Methods

Data preprocessing and analysis were performed using the Searchmight Toolbox (35) for Matlab, AFNI functions (36), and custom scripts. Further methodological details are provided in *Supporting Information*. There, we describe scan parameters, participants, stimuli, experimental procedure, data analyses, and additional results. All participants gave informed consent in accordance with the guidelines of the Committee on the Use of Human Subjects at Harvard University.

ACKNOWLEDGMENTS. We thank Fiery Cushman, Steven Pinker, Alfonso Caramazza, Susan Carey, and Patrick Mair for helpful comments. We thank Anita Murrell, Sarah Coughlon, Frantisek Butora, and Rebecca Fine for research assistance. This work was supported by a National Science Foundation Graduate Research Fellowship (to S.M.F.).

1. Frege G (1976) *Logische Untersuchungen* (Vandenhoeck und Ruprecht, Göttingen), 2, Erg. Aufl. Ed.
2. Pinker S (1994) *The Language Instinct* (Morrow, New York), 1st Ed.
3. Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1-2):3–71.
4. Smolensky P (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif Intell* 46(1-2):159–216.
5. Marcus GF (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (MIT, Cambridge, MA).
6. Shastri L, Ajjanagadde V (1993) From simple associations to systematic reasoning—a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behav Brain Sci* 16(3):417–451.
7. von der Malsburg C (1999) The what and why of binding: The modeler's perspective. *Neuron* 24(1):95–104, 111–125.
8. Doumas LA, Hummel JE, Sandhofer CM (2008) A theory of the discovery and predication of relational concepts. *Psychol Rev* 115(1):1–43.
9. O'Reilly RC, Busby RS (2002) Generalizable relational binding from coarse coded distributed representations. *Adv Neural Inf Process Syst* 1:75–82.
10. Kriete T, Noelle DC, Cohen JD, O'Reilly RC (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc Natl Acad Sci USA* 110(41):16390–16395.
11. Wu DH, Waller S, Chatterjee A (2007) The functional neuroanatomy of thematic role and locative relational knowledge. *J Cogn Neurosci* 19(9):1542–1555.
12. Pallier C, Devauchelle AD, Dehaene S (2011) Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci USA* 108(6):2522–2527.
13. Fedorenko E, Behr MK, Kanwisher N (2011) Functional specificity for high-level linguistic processing in the human brain. *Proc Natl Acad Sci USA* 108(39):16428–16433.
14. Hagoort P, Hald L, Bastiaansen M, Petersson KM (2004) Integration of word meaning and world knowledge in language comprehension. *Science* 304(5669):438–441.
15. Friederici AD, Rüschemeyer SA, Hahne A, Fiebach CJ (2003) The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cereb Cortex* 13(2):170–177.
16. Vandenberghe R, Nobre AC, Price CJ (2002) The response of left temporal cortex to sentences. *J Cogn Neurosci* 14(4):550–560.
17. Bemis DK, Pylkkänen L (2011) Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J Neurosci* 31(8):2801–2814.
18. Baron SG, Thompson-Schill SL, Weber M, Osherson D (2010) An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cogn Neurosci* 1(1):44–51.
19. Baron SG, Osherson D (2011) Evidence for conceptual combination in the left anterior temporal lobe. *Neuroimage* 55(4):1847–1852.
20. Humphries C, Binder JR, Medler DA, Liebenthal E (2006) Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J Cogn Neurosci* 18(4):665–679.
21. Devauchelle AD, Oppenheim C, Rizzi L, Dehaene S, Pallier C (2009) Sentence syntax and content in the human temporal lobe: An fMRI adaptation study in auditory and visual modalities. *J Cogn Neurosci* 21(5):1000–1012.
22. Rogalsky C, Hickok G (2009) Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cereb Cortex* 19(4):786–796.
23. Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539):2425–2430.
24. Mitchell TM, et al. (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–1195.
25. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10(9):424–430.
26. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103(10):3863–3868.
27. Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 51(6):1173–1182.
28. Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48(2):175–187.
29. Mackinnon DP, Lockwood CM, Williams J (2004) Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behav Res* 39(1):99–128.
30. Selig JP, Preacher KJ (2008) Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects. Available at quantpsy.org/. Accessed May 24, 2015.
31. Skeide MA, Brauer J, Friederici AD (2014) Syntax gradually segregates from semantics in the developing brain. *Neuroimage* 100:106–111.
32. Kuperberg GR, et al. (2000) Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: An fMRI study. *J Cogn Neurosci* 12(2):321–341.
33. Fodor JA (1975) *The Language of Thought* (Harvard Univ Press, Cambridge, MA).
34. Plate TA (1995) Holographic reduced representations. *IEEE Trans Neural Netw* 6(3):623–641.
35. Pereira F, Botvinick M (2011) Information mapping with pattern classifiers: A comparative study. *Neuroimage* 56(2):476–496.
36. Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29(3):162–173.
37. Kutas M, Hillyard SA (1980) Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207(4427):203–205.
38. Friston KJ (2011) Functional and effective connectivity: A review. *Brain Connect* 1(1):13–36.
39. Kriegeskorte N, et al. (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60(6):1126–1141.
40. Chiu YC, Esterman MS, Gmeindl L, Yantis S (2012) Tracking cognitive fluctuations with multivoxel pattern time course (MVPTC) analysis. *Neuropsychologia* 50(4):479–486.
41. Coutanche MN, Thompson-Schill SL (2013) Informational connectivity: Identifying synchronized discriminability of multi-voxel patterns across the brain. *Front Hum Neurosci* 7:15.
42. Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN (2008) Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron* 59(6):1037–1050.
43. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* 4:e32.
44. Duda RO, Hart PE, Stork DG (2001) *Pattern Classification* (Wiley, New York), 2nd Ed.

Supporting Information

Frankland and Greene 10.1073/pnas.1421236112

Experiment 1: Supporting Materials, Methods, and Results

Subjects. Eighteen self-reported right-handed subjects, 10 male, from the Harvard University community participated in experiment 1 for payment (aged 19–34). All subjects were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with Harvard University's institutional review board. Data from two subjects were not analyzed due to failure to properly complete the experiment. No subjects exhibited excessive head motion. Experiment 1 analyses used the remaining 16 subjects.

Stimuli and Experimental Procedure. The sentences used in experiment 1 are listed in Table S1. All sentences contained transitive verbs, and described contact events (e.g., hit) involving two entities. Four of the mirror image proposition pairs described situations low in negative emotional valence and moral wrongness, as rated by an independent group of subjects ($n = 33$) recruited through Amazon Mechanical Turk (https://www.mturk.com/). Two of these four pairs involved inanimate entities (e.g., "the truck hit the ball"/"the ball hit the truck"), and two involved animate entities (e.g., "the girl touched the grandmother"/"the grandmother touched the girl"). Both animate and inanimate entities were included to better localize regions encoding domain-general structure-dependent meaning. Information regarding the frequency of occurrence of the used sentences is provided in *Stimulus Frequency and lmSTC* below. The second set of mirror image proposition pairs were judged to be asymmetrically emotionally evocative and asymmetrically morally wrong, depending on which entity was described as performing the action. (e.g., "the grandfather kicked the baby" worse than "the baby kicked the grandfather") ($P < 0.01$, for all analyses). For these two items, one proposition was therefore expected to produce a downstream affective response that its mirror image would not produce, either in kind or magnitude, and hence were used in our pattern-based effective connectivity (PBEC) analysis.

We used a slow event-related design in which sentences were presented visually for 2.5 s, followed by 7.5 s of fixation. Pseudorandom stimulus presentation lists were generated according to the following constraints: Each proposition was presented twice within each run, and neither the same proposition, nor a proposition and its mirror image could be presented successively, to avoid any overlap of the hemodynamic response for the to-be-discriminated items. The experimental session consisted of 13 scan runs, resulting in 26 presentations of a given proposition over the course of the experiment. For one participant, only 9 of the 13 runs were available for analysis due to technical problems.

Whether the proposition was presented in the active or passive voice on a given trial was randomly determined. Active and passive versions of the same proposition were treated identically for all analyses. Three strings of nonwords were also presented to subjects in each run, but were not analyzed. On one-third of the trials, questions were presented following the fixation period. These consisted of questions about the agent of the immediately preceding proposition (e.g., "Did the ball hit something?"), questions about the patient ("Did the ball get hit by something?"), or prompts to rate "how morally bad" the event was on a scale of 1–5, with one being "not bad at all" and 5 being "very bad." Fifty percent of the comprehension questions had affirmative answers. The subjects' responses were signaled using a right-hand button box. Which question was presented on a given trial was randomly determined, as were the particular trials that were followed by comprehension questions. These were included simply to promote subject engagement and were not analyzed.

Experiment 1: Classification Analyses

Whole-Brain Searchlight Mapping. We used a whole-brain searchlight procedure (26) to determine whether any brain regions reliably contained information about the meaning of the presented sentences across subjects. Following the approach of Mitchell et al. (24), we averaged over the temporal interval from 2.5 to 10 s following stimulus onset to create a single image for each trial. The two presentations of each proposition for a given run were then averaged to create a single image per proposition, per run. All experiment 1 analyses were performed on these averaged images.

We conducted our searchlight analyses using the Searchmight Toolbox (35). A cube with a 2-voxel (6-mm) radius was centered at each voxel, and a linear discriminant classifier with a shrinkage estimate of the shared population covariance matrix was used to probe the surrounding region for informational content. Nonedge neighborhoods contained 124 voxels. Pairwise classifiers were separately trained for each of the four mirror image proposition pairs. For every pair, performance at a given location was assessed by iteratively holding out each run as test data, training the pairwise classifier on 12 of 13 runs, testing on the held-out run, and averaging performance across the 13 cross-validation folds. This resulted in a single whole-brain accuracy map per mirror image proposition pair. These four pair-level maps were then averaged to create one map across pairs for each subject, with the aim of identifying regions that consistently contained information across the four mirror image pairs.

These maps were then spatially normalized to Talairach space for group-level statistical analyses. Given that all comparisons were pairwise, we assume a mean of 0.5 for the null distribution, and test the directional hypothesis that a given region contains information across subjects by performing a one-tailed t test against 0.5 on the set of accuracy maps. Control of the whole-brain familywise error rate was obtained through a combination of voxelwise thresholding and cluster extent. These corrected P values were obtained through Monte Carlo simulations in AFNI (36). Such simulations empirically estimate the probability of obtaining clusters of statistically significant values, given that the data contain only noise. To estimate the smoothness of the noise, we conducted an analysis that randomly permuted the sentence labels for each subject and mimicked the individual and group procedures above to obtain a group "noise-only" map. We thus used the actual data with the same analytic operations to estimate the smoothness of the noise. The resulting spatial smoothness of this noise-only map was found to be $X = 6.2$, $Y = 6.34$, and $Z = 6.3$ mm. These dimensions were thus used in the Monte Carlo simulations to estimate the probability of obtaining significant clusters across the whole-brain volume, given only noise.

We first chose a voxelwise threshold of $P < 0.005$ and a corrected threshold of $P < 0.05$, but no statistically significant clusters survived this threshold. Given that our chief aim was simply to localize candidate regions for our connectivity analysis and for analysis in experiment 2, we performed the same analysis at three lower voxelwise thresholds: $P < 0.01$, $P < 0.02$, and $P < 0.05$. At a threshold of $P < 0.02$, two regions of interest (ROIs) were found to be highly statistically significant [left mid-superior temporal cortex (lmSTC), 123 voxels, clusterwise $P < 0.0001$: right extreme capsule/insula, 108 voxels, $P < 0.001$ clusterwise]. These effects easily survive Bonferroni correction for the use of four voxelwise thresholds. Thus, although the observed voxelwise effects are relatively small, requiring a liberal voxelwise threshold for detection, they yield a clusterwise

effect in lmSTC that is highly robust. The large extent of this ROI makes it unlikely that it consists of a single functional unit, however. With this in mind, experiment 2 probes the informational content of one of these two regions.

The first discovered ROI was located in the lmSTC. It begins in the inferiormost part of the parietal cortex, extends through the midportion of the superior temporal gyrus and superior temporal sulcus and terminates in the middle temporal gyrus (Fig. S1). The second ROI is centered on the right posterior insula/extreme capsule fiber bundle. It encompasses parts of the posterior insula, claustrum, putamen, and medial-superior temporal lobe ($P < 0.001$, 108 voxels) (Fig. S1).

Post Hoc ROI Analyses. As our principal searchlight results from experiment 1 were obtained by averaging classification accuracy across four pairs, it is possible that the significance of the results owes to high accuracy on some subset of the pairs, with chance performance on the remaining pairs. To evaluate this possibility, we trained and tested linear discriminant classifiers with a shrinkage estimate of the covariance matrix separately on each pair in lmSTC to evaluate post hoc which mirror image proposition pairs were driving our results. Table S2 shows the results of these classifications by mirror image proposition pair and ROI. A repeated-measures ANOVA revealed no significant differences in classification accuracy across the six pairs for either lmSTC [$F_{(5,75)} = 0.4$, $P = 0.84$] or the right insula/extreme capsule ROI [$F_{(5,75)} = 0.15$, $P = 0.98$]. We find fairly consistent levels of classification accuracy in both ROIs, suggesting these regions are not driven by idiosyncrasies of the particular pairs, or only by the animate or inanimate proposition pairs. Instead, the pattern of results is consistent with both lmSTC and the right posterior insula/extreme capsule encoding domain-general information about "who did what to whom."

If the regions discovered in the searchlight analysis do represent structure-dependent meaning, then they should facilitate classification of nonmirrored propositions as well. For example, they should be able to distinguish "the truck hit the ball" from "the father pulled the child." Although these nonmirrored pairs are not well matched, and one would expect many other brain regions to be able to perform this classification (e.g., regions that encode the semantic/phonological content of the nouns and verbs), this analysis nevertheless serves as a "sanity check" on the ROIs localized using the searchlight analysis. To ensure that our ROIs could also discriminate nonmirrored pairs, pairwise classifiers were separately used for the 24 nonmirrored comparisons that could be generated from the four mirror image proposition pairs. This analysis was performed using data from the lmSTC ROI and the right posterior insula/extreme capsule ROI separately. These 24 classification accuracy statistics were then averaged for each ROI and submitted to a one-tailed t test against 0.5. Of the two ROIs able to discriminate within mirror image proposition pairs, only the lmSTC ROI was able to reliably discriminate nonreversed pairings as well [$t_{(15)} = 4.06$, $P = 0.005$]. The right extreme capsule/insula ROI trended in this direction, but its results were not statistically significant [$t_{(15)} = 1.45$, $P < 0.09$]. This failure to robustly classify non-mirror image proposition pairs casts doubt on the possibility that the right posterior insula encodes complex, structured semantic representations. Taken in conjunction with other null results pertaining to this ROI, described in *PBEC Analyses* below, we chose to focus on lmSTC in experiment 2.

Stimulus Frequency and lmSTC. The sentences used were chosen partly because of their relative infrequency in English, ensuring that subjects would not recognize the sentences as familiar units, and would not have strong expectations about which entity is most likely to be assigned to which role. (We did not use the familiar sentences "Dog bites man" and "Man bites dog" in our experiments for these reasons.) See refs. 14, 15, and 37 for literature on such violation of expectations, and the resulting N400 response in electrophysiology. For the used sentences, no active-voice construction was present in the Google 5-gram text corpus (https://catalog.ldc.upenn.edu) or the Google Books ngram corpus (https://books.google.com/ngrams) as of August 2011. This strongly suggests that frequency differences within pairs were not responsible for the observed classification performance. Moreover, as both propositions within a pair were composed of the same words, the pairwise summed word frequency was necessarily identical.

It remains possible, however, that the frequency of various higher-order parts of the sentences could differ, even if the entire sentences that contain these parts are matched. For example, the construction "the father pulled" may be more frequent than the construction "the child pulled," which could facilitate pairwise discrimination of "the father pulled the child" and "the child pulled the father." If frequency statistics were driving differences in the observed patterns of activity, then we would expect the region to carry information about these statistics across propositions. To address this possibility, we attempted to predict various frequency statistics of the sentences from lmSTC's patterns of activity.

We trained separate regression models to predict various frequency statistics pulled from the Google Ngram corpus (https://catalog.ldc.upenn.edu) and simple Google web search (www.google.com). These statistics included the frequency of the agent/verb combinations in the active voice (e.g., "the father pulled" and "father pulled"), verb/patient combinations (e.g., "pulled the child), and the mean, minimum, and maximum of these statistics for a given proposition. Each proposition was first described by a log transformation of the relevant frequency statistic, and a support vector regression (SVR) model was trained to predict the continuous value of that statistic from the pattern of activity in lmSTC.

To evaluate the models, we trained the SVR model on $N - 2$ trials and attempted to predict the frequencies of two held-out observations. The absolute value of the difference between the target frequencies and those predicted by the SVR model was compared for both the correct mapping and the incorrect mapping. If the sum of the two correct mappings had a lower absolute difference than the sum of the two incorrect mappings, the model was determined to have been correct. This procedure was repeated using different frequency metrics, and the results are shown in Fig. S2. We consistently found no information about any frequency statistics to be available in lmSTC. Although these results may appear to be below a priori chance levels of 50%, they are within the range that we discover randomly permuting the labels and running the analysis 1,000 times for each subject. The mean accuracy using randomly permuted data are 49.6%, with 90% of the mean accuracies falling between 45.4% and 53.5%. Across frequency statistics, we find a mean accuracy of 47.5%, which falls within the expected range. We also note that our frequency measures are correlated, making it less surprising that multiple frequency measures lead to below-chance performance, given that any do. It is therefore unlikely that our results owe to systematic frequency differences between pairs: Rather, they appear to reflect the structured, semantic content of sentences.

PBEC Analyses

Background and Motivation. The identification of representational content in the human brain has benefited from the development of multivoxel pattern analysis (MVPA) (23–26). Rather than asking whether the magnitude of the blood oxygen level-dependent (BOLD) response in a single voxel or brain region is predicted by the presence or absence of a psychological operation, researchers now routinely pool information across sets of voxels to ask whether and where distributed patterns of activity track variation in psychological content. MVPA has been productively applied to domains lacking the differential engagement

of psychological processes expected to generate uniformly greater neural activity over a brain region (when sampled at the spatial resolution available to contemporary neuroimaging), but in which some representational content nevertheless varies over time. Although this ability makes MVPA particularly well suited to our current aims, its increased power is not completely without cost, as its heightened detection sensitivity makes it more susceptible to subtle confounds. It is therefore particularly important to establish that information detected by the pattern classifier actually reflects the psychological processes or representations of interest.

To address this potential concern, we used the following reasoning: If the patterns identified reflect neural representations of psychological content, then one would expect the pattern instantiated on a given trial to modulate downstream responses that depend on that content. Given that this logic is broadly consistent with the logic underlying traditional effective connectivity analyses (38), we call the present analysis a PBEC analysis. As with conventional effective connectivity analyses, the aim is to establish the functional influence of one neuronal population on another.

Several groups have recently integrated MVPA and functional connectivity analyses, devising ways to determine whether various neural structures share similar representational profiles (39), and whether patterns in one region correlate with univariate responses (40) and patterns (41) elsewhere in the brain. The present analysis extends this integration of MVPA and connectivity analyses to model cases in which the patterns of activity in one region are thought to drive the functional state of another using mediation analyses (27). Such tests are widely used in the social sciences (27) and have been previously applied to fMRI data (42). Tests for mediation assess whether the effect of a predictor variable on an outcome variable is either partly or wholly carried by an intervening, or mediating, variable.

In the present context, if (*i*) the pattern of activity instantiated across a region reflects the neural representations of interest, and (*ii*) those representations are hypothesized to drive an independent response, then (*iii*) that pattern may mediate the effect of the stimulus on this downstream response. Here, we used mirror image propositions that differ in their affective significance, such as "the grandfather kicked the baby" and "the baby kicked the grandfather," and sought to determine whether their associated patterns, instantiated across lmSTC, mediate the relationship between the sentences presented and the consequent affective responses to these sentence's meaning. Such a finding would provide evidence that the identified patterns do indeed reflect neural representations of these sentences' meaning.

To succeed, the mediating variable (here, the pattern of activity in lmSTC) must explain unique variance in the downstream response over and above variance in that response explained by the stimulus. This is because, to the extent that there is variability or "error" in this causal process, the more proximate mediating variable (here, the pattern of neural activity) should explain unique variance in the response, in virtue of being a channel through which the direct effect (here, the effect of the proposition presented on the amygdala's activity level) is carried. The detailed procedure for quantifying the pattern of activity in lmSTC and testing the mediation hypothesis is specified below.

PBEC Procedure. For a binary classification problem, such as discriminating patterns evoked by "the grandfather kicked the baby" and "the baby kicked the grandfather," the training procedure establishes a hyperplane that divides the feature space (in this case, a voxel-activity space) into two regions. Here, one region is associated with the characteristic pattern of one proposition and one with the characteristic pattern of the other (Fig. 1). Each trial's multivoxel BOLD response then lies at some distance and direction from this classification hyperplane in one of the two regions of the space.

We used these trial-by-trial "signed distances" as measures of the representational content of lmSTC on a given trial, as they carry information both about the classifier's decision (the sign, corresponding to the side of the hyperplane and region of the space) and, roughly, its "confidence" (the absolute value of the distance). This effectively reduces the dimensionality of the region from the number of voxels in the ROI to one. Here, that one variable summarizes the informational content of psychological interest contained by the entire region. In the current coding scheme, good instances of the affectively salient proposition ("the grandfather kicked the baby") will have large positive distances, good instances of the affectively neutral proposition ("the baby kicked the grandfather") will have large negative distances, and ambiguous instances will have distances near zero. We can then ask whether these distance variables predict responses elsewhere in the brain. These signed distance variables were obtained and used as follows.

First, linear classification functions were learned separately for the two mirror image proposition pairs using a leave-one-out procedure.

The classification function for each novel test exemplar is then given by the following:

$$g(x) = w^T x + w_0,$$

where x is the vector of voxel intensities for the current test trial, and ^{T} denotes vector transposition. The voxel-weight vector w for each cross-validation iteration was obtained as follows:

$$w = \Sigma^{-1}(m_1 - m_2),$$

where m_1 and m_2 are the vectors of class-specific mean voxel intensities across lmSTC. m_1 is the affectively salient ("the grandfather kicked the baby") mean vector, and m_2 is the affectively neutral ("the baby kicked the grandfather") mean vector. Σ^{-1} was a shrinkage estimate of the population covariance matrix shared between all stimulus classes (43).

The constant term was determined as follows:

$$w_0 = -\frac{1}{2}\left(m_{1(group)} - m_{2(group)}\right).$$

The weight vector w determines the direction through the feature space, whereas the constant term determines the location of the hyperplane relative to the origin.

Here, the m_1 and m_2 terms for each subject were the means of the respective class along the projection for the remaining 15 subjects. We found these group-level mean estimates to yield more reliable predictions in the connectivity analysis than using an individual subject's data [$t_{(24)} = 3.93$, $P = 0.001$, vs. $t_{(24)} = 1.28$, $P = 0.22$]. These results survive correction for multiple comparisons for these two ways of obtaining w_0.

Finally, the signed distance of an observation from the hyperplane was computed as in ref. 44:

$$d = \frac{\left(w^T x + w_0\right)}{||w||},$$

where $||w||$ is the Euclidean norm of the weight vector, representing the distance from the origin to w.

This distance provides a trial-by-trial measure of the representational content of lmSTC, which can be used as a predictor variable in subsequent connectivity analyses. In the present case, given that the affectively significant propositions occupied the positive region of the space, and the affectively neutral regions occupied the negative region of the space, we predict a positive statistical relationship between trial-by-trial signed distance in

lmSTC and the magnitude of the mean signal level across the left amygdala ROI (Fig. 1).

We expect the magnitude of the distance variable to explain unique variance over and above the sign of the distance variable because the probability that the classifier is correct should increase with an increase in distance. This assumption was borne out empirically. Across all pairs, the absolute value of distance from the hyperplane predicts classifier performance ($P < 0.001$). For example, those trials in which the classifier is relatively confident that the content of the stimulus is "the grandfather kicked the baby" relative to "the baby kicked the grandfather," are in fact more likely to be correct (and vice versa).

We did not, however, expect to find a perfectly linear relationship between the signed distance from the classification boundary and the amygdala response for a number of reasons: First, we do not expect meaningful variation in the prediction of the amygdala response as a function of distance on the "the baby kicked the grandfather" side of the hyperplane. We see no reason that "good" instances of the affectively neutral class should be less likely to elicit an amygdala response than bad instances of that class, as a linear model would predict. Furthermore, although we would expect good instances of the affectively salient proposition to be more likely to elicit an amygdala response than affectively neutral trials, or trials about which the classifier is uncertain, it is not obvious that this relationship should continue indefinitely in a linear manner, without saturation. We therefore sought a way to incorporate the continuous information provided by an observation's distance from the classification boundary, without confining ourselves to a simple linear predictive model. We thus chose transform the signed distance from the hyperplane with a sigmoidal function, which (conceptually) applies both a threshold (a point below which we would not expect an amygdala response) and a point of saturation (a point above which we not expect increasing distance to predict an increased likelihood of amygdala response).

The precise shape of the sigmoid is controlled by two free parameters: one affecting the center of the function ($p2$ below), and the other ($p1$) affecting its slope:

$$s = \frac{1}{1 + e^{p1(p2-d)}}.$$

Given that we did not have a priori quantitative predictions for the precise shape of the function, we allowed the value of the two parameters to be determined empirically through cross-validation. For each subject, the 2D parameter space (center X slope) was searched with that subject's data removed. Coefficients for the regression of amygdala activity on the transformed lmSTC signed distance variable were obtained for each combination of the center and slope parameters. The parameter combination yielding the best prediction, defined as the greatest mean β value, on the remaining subjects was then used for the held-out subject. The parameter combination selected was stable across cross-validation iterations.

The heat map in Fig. S3 visualizes the average regression performance for various sigmoids by averaging the search results across 16 cross-validation iterations (one holding each subject out). These search results provide information about the relationship between the pattern in lmSTC and the amygdala's response. For simplicity, if we conceptualize the amygdala response as a binary variable, the center parameter of the sigmoid defines the point at which the probability of a response is equally likely to the probability of a nonresponse. We see from the heat map in Fig. S3 that the optimal center of the sigmoid is shifted to the right of the classification boundary, in the positive region of the space. The observed positive shift of the center parameter relative to the hyperplane is explicable under the assumption that the probability of the amygdala's not responding given that the stimulus is "the grandfather kicked the baby," is greater than the probability of the amygdala's responding given that the stimulus was "the baby kicked the grandfather." This assumption is reasonable given that each proposition is encountered repeatedly over the course of the experiment, potentially attenuating the subject's affective responses over repeated presentations. This would make failures to respond to the affectively salient proposition more likely than "false positives" of the amygdala to the affectively neutral proposition, leading to a positive shift of the optimal center (as the point of equiprobability) for the sigmoid relative to the classification hyperplane. This method may hold more general promise for testing different quantitative models of the functional dependence between brain regions.

Finally, we asked whether the signed distance of an observation from the classification hyperplane in lmSTC mediates the relationship between the stimulus and the mean level of activity in the amygdala. To satisfy conventional criteria for mediation, it is necessary that the pattern of activity in lmSTC explain variance in the amygdala's response over and above the variance explained by the identity of the stimulus presented (here, which proposition the subject read). We therefore included a binary regressor coding the content of the presented proposition as a covariate of no interest. We also included a regressor for the mean signal level across the entire lmSTC ROI as a second covariate of no interest, to preclude the possibility that any observed mediation was due solely to aggregate functional coupling between the regions. We obtained standardized coefficients for the regression of amygdala signal on signed distance separately for all subjects, and we used a one-sample t test to evaluate whether this coefficient was reliably nonzero across subjects. Finally, we used the classical Sobel test (27), as well as Monte Carlo simulation (29, 30), to assess the significance of the indirect effect $a*b$ on the dependent variable.

The Sobel statistic is computed as follows:

$$z = \frac{ab}{\sqrt{s_a^2 b^2 + s_b^2 a^2 + s_a^2 s_b^2}}.$$

In the present experiment, a is the ordinary least-squares regression coefficient of distance from the hyperplane on the category label, and b is the regression of mean left amygdala on this distance, controlling for the category label of the stimulus, and the mean signal level across the ROI. The s terms in the denominator are the SEs for the a and b coefficients, respectively.

As reported in the main text, we found this indirect effect to be statistically significant ($z = 2.47$, $P = 0.013$) using the Sobel test. Given that the Sobel test is unreliable for small sample sizes (29), we also used Monte Carlo simulation to estimate confidence intervals for the indirect effect. This procedure yielded results comparable to the Sobel test ($P < 0.01$ for the indirect effect). The above analyses were repeated with the right posterior insula/extreme capsule ROI as the mediator. In contrast to lmSTC, all assessments of the pattern of activity in the right insula ROI as the mediating variable were nonsignificant ($P > 0.15$).

What is the source of the trial-by-trial variability driving these results? One possibility is that subjects are habituating to the sentences over the course of the experiment, resulting in a systematic decrease in the distance from the hyperplane over time. We find a nonsignificant negative relationship between trial number and absolute distance from the hyerplane ($r = -0.05$) for the two mirror image proposition pairs used in the PBEC analysis. Although very small, this effect trends toward significance when the correlation coefficients are estimated separately for each subject, and then pooled across subjects [$t_{(14)} = 1.85$, $P = 0.08$]. We see similar trends when we bin the data by run, rather than maintaining the temporal ordering of each trial [mean $r = -0.16$, $t_{(14)} = 1.79$, $P = 0.095$]. The

data therefore show some weak signs of habituation; however, from these analyses, we find no strong evidence that distance from the hyperplane is decreasing systematically over time. It therefore seems as though there are additional sources of trial-by-trial variability. These may be transient fluctuations in participant attention from trial to trial. Or they may be more systematic sources of variability, such as that introduced in passive sentences: Passive sentences require additional syntactic operations to recover who did what to whom. This additional computation may introduce variability in the semantic representation in lmSTC and therein also attenuate affective responses.

Whatever the source, we find evidence that the relationship between lmSTC and the amygdala is specific and systematic; when the pattern corresponding to "the grandfather kicks the baby" rather than "the baby kicked the grandfather" is instantiated, we see an increased amygdala response controlling for the factors detailed above.

Although we believe the emotional response is best reflected by the mean response amplitude of the entire left amygdala ROI, we also asked whether the pattern in lmSTC significantly predicts the pattern in the amygdala. To evaluate this possibility, we trained a pattern classifier to differentiate emotionally salient mirror image proposition pairs and used this classifier's distance from the hyperplane as the outcome instead of the mean amygdala response. We found a significant correlation between distance from the hyperplane in lmSTC and distance from the hyperplane in the amygdala, when trained to classify the stimuli, controlling for the stimulus, and the mean level of activity in lmSTC [$t_{(15)} = 2.41$, $P = 0.028$]. As with the mean amplitude, the pattern in the amygdala does not predict the pattern in lmSTC [$t_{(15)} = 1.1$, $P = 0.29$], controlling for the stimulus. Using the pattern in the amygdala, rather than the mean signal, as the dependent measure, we find a trending, but nonsignificant mediation effect ($z = 1.88$, $P = 0.06$). These results are therefore also consistent with the proposed causal model, although the effects are smaller than when using the mean regional response in the amygdala as the dependent variable.

Experiment 2: Supporting Information

Subjects. Thirty-four self-reported right-handed members of the Harvard community participated for payment (aged 18–35). We used the same subject inclusion criteria as in experiment 1. One subject's data were not analyzed to due to failure to properly complete the experiment. Six subjects' data were excluded before analysis due to answering less than 75% of the comprehension questions correctly. Two subjects were excluded for exhibiting excessive head motion, defined as greater than 3 SDs above the mean. Data from the remaining 25 subjects were included in all experiment 2 analyses.

Stimuli and Experimental Procedure. The sentences for experiment 2 were generated using four nouns ("man," "girl," "dog," and "cat") and five transitive verbs ("chased," "scratched," "blocked," "approached," and "bumped") to create every possible agent–verb–patient combination, with the exception of combinations using the same noun twice (e.g., "the dog chased the dog") yielding 60 ($4 \times 5 \times 3$) unique propositions. These particular verbs were chosen because they permit plausible agent–verb–patient combinations using the above nouns, and are comparable in their frequency of occurrence.

Experiment 2 consisted of six scan runs. Each proposition was presented once per run, and six times in total. Whether a proposition was presented in the active or passive voice on a given trial was randomly determined. As in experiment 1, sentences were visually presented for 2.5 s followed by 7.5 s of fixation. A comprehension question was presented following the fixation period on one-third of the trials. These questions were of the form "Did the dog chase something?" or "Was the dog chased by something?" and 50% had affirmative answers.

General Searchlight Procedure. All searchlight analyses for experiment 2 were confined to the lmSTC. The searched area was formed by dilating the group-level lmSTC ROI discovered in experiment 1 6 mm so as to encompass all of the mid and posterior regions of the left superior temporal gyrus, superior temporal sulcus, and middle temporal gyrus. The resulting ROI contained 6,882 1.5-mm^3 voxels, (center, −54, 23, 3). Fig. 2*B* shows the extent of the searched region for experiment 2. This mask was warped from Talairach space to each subject's native space, and all classification analyses were conducted in the subject's native space. As in experiment 1, all searchlight analyses were implemented in the Searchmight Toolbox (35) and used a linear classifier with a shrinkage estimate of the covariance matrix. Local voxel-neighborhoods were defined using a 3-mm (two-voxel) radius within the lmSTC mask, entailing that nonedge neighborhoods again contained 124 voxels.

Agent and Patient Decoding Procedure. For our principal analyses, we searched lmSTC for patterns of activity encoding information about the identity of the agent and patient that generalizes across verbs. Agent and patient classifications were performed using separate classifiers, iteratively using data from local voxel-neighborhoods to make four-way decisions regarding the noun occupying the agent or patient role on a given trial (man? girl? dog? cat?). As in experiment 1, active and passive versions of the same proposition were considered identical for the purposes of these analyses. To train and test the classifier, we used a fivefold cross-validation procedure defined over the five verbs ("chased," "scratched," "blocked," "approached," "bumped"). For a given iteration, all data generated by one of the five verbs was removed, and classifiers were trained to identify the noun occupying the agent or patient role on that trial, using the data from the remaining four verbs. The classifiers were then tested using the patterns generated by the held-out verb.

Classification accuracies for each subject were averaged across cross-validation folds, and the mean accuracy was assigned to the center voxel of the search volume. These individual-level accuracy maps were then smoothed with a 3-mm FWHM kernel, warped to Talairach space, and the group of subjects' maps was submitted to a directional one-sample *t* test against 0.25 to determine whether any regions reliably encoded information about the identity of the agent or patient across subjects. We used Monte Carlo simulation to determine the probability of obtaining significant clusters given that the data contained only noise. We used a voxelwise threshold of $P < 0.005$, and a corrected threshold of $P < 0.05$ to identify regions exhibiting statistically significant effects. As in experiment 1, we estimated the smoothness of the data by conducting the same classification and aggregation procedures with randomly permuted labels. The obtained smoothness parameters for each analysis, as well as information about the voxel clusters found to be significant by the Monte Carlo simulation, are presented in Table S3.

Verb Decoding Procedure. We performed an additional searchlight analysis probing lmSTC for subregions that contained information about the trial-by-trial identity of the sentence's verb. This task required a five-way decision regarding the identity of the verb ("chased," "scratched," "blocked," "approached," "bumped"). Here, the cross-validation folds were defined over the six scanning runs. The classifiers were trained using data from five of six runs, and were then asked to identify the verb present for trials from the remaining run. The verb classifiers were thus only required to generalize to new tokens of previously encountered combinations, rather than wholly new combinations. Individual accuracy maps were averaged as in prior analyses, smoothed with a 3-mm FWHM kernel, warped to Talairach space, and submitted to a one-tailed *t* test against 0.2, as chance performance for the five-way verb classification was 20%. Corrected *P* values were obtained

in the same manner as in the above searchlight analyses. Results are visualized in Fig. S4, and more information is provided about the analysis and results in Table S3.

Post Hoc ROI Analyses.

Representational specificity of subregions. For all post hoc analyses, we used a leave-one-subject-out cross-validation procedure to localize regions of interest without biasing the analyses. Specifically, we iteratively conducted group-level t tests on the search maps for 24 of 25 subjects to identify clusters of informative voxels with each subject's data removed. All such regions were localized using a voxelwise threshold of $P < 0.005$ and a minimum cluster size of 50 mm^3, unless otherwise specified. These voxel clusters defined the exact ROI for the held-out subject for the post hoc test of interest. For any post hoc analysis, the exact ROI queried could thus vary slightly from subject to subject.

We first assessed the representational specificity of the agent and patient subregions identified by the searchlight analysis, as described in the main text. After localizing clusters using the leave-one-subject-out cross-validation method described above, we averaged the held-out subject's agent and patient search results across the voxels contained by the subregions. This produced one average accuracy statistic for each of the agent and patient identification problems, in each of the three subregions (two agent, one patient), for each subject. We performed repeated-measures ANOVAs to test for a subregion by content interaction for the anterior agent and patient regions, and paired t tests to test for simple effects of identification accuracy within each of these three ROIs. Results of these analyses are reported and plotted in the main text (Fig. 3). The average classification accuracies for the voxels comprising these ROIS were small, but reliably above chance.

In addition to assessing significance using the analytic P values of the F and T distributions, we also performed permutation analyses. To do so, we randomly scrambled the labels of the classification accuracies (agent/patient), holding the subject and region constant. For each subject, this analysis randomly assigns "agent" classification accuracies to either the "agent" or "patient" region, and with different randomizations across subjects, and likewise for patient classification accuracies. This particular permutation approach has a number of desirable properties: First, it preserves the within-subject structure of the data; accuracies obtained from a given participant are shuffled within that participant. Second, it preserves the across-subjects structure of the data, because the regions used for the permutation analysis were identified using $N - 1$ subjects' data, just as in the original analysis. We performed this procedure 100,000 times to generate an empirical null distribution. The empirically observed probabilities very closely correspond to those derived analytically from the F and T distributions in all cases but one: Classification of the agent in the anterior agent ROI (as defined using independent subjects) is statistically significant when assessed using the analytic t distribution ($P = 0.02$), but only marginally significant when using the permutation test ($P = 0.055$). (However, see a significant replication of this finding in experiment 3 below.)

Next, because we failed to find a significant difference between agent identification accuracy and patient identification accuracy within the agent subregions, we asked whether these regions might simply be encoding information about the semantic content of the nouns without respect to semantic role. To evaluate this possibility, we performed two additional searchlight analysis within lmSTC. In the first, classifiers were trained to determine whether a given noun was present in the sentence, regardless of whether it was the agent or patient. For example, "The dog chased the man" and "The cat scratched the dog" were coded identically, given that they both contain an occurrence of "dog." The classifier's task was to determine whether or not "dog" was present in the sentence. This procedure was repeated for each of the four nouns, and the results were averaged to obtain one classification accuracy statistic for each voxel neighborhood.

To directly compare such "role-neutral" identification functions to the "role-based" identification functions we have reported thus far, we performed a searchlight analysis in which classifiers made pairwise, role-based decisions ("man" as agent vs. "cat" as agent). It was necessary to perform this pairwise role-based classification analysis given that chance levels of performance differed between role-neutral and our primary role-based classification functions: Chance performance for "role-neutral" classifiers was 50%, whereas chance performance for "role-based" classifiers was 25%. However, because each word appears in both agent and patient roles, these role-neutral classification functions naturally have twice as much data available to train and test the classifier. To equate the amount of the data, we randomly sampled 50% of each subject's data for use in the role-neutral classification task to equate the amount of data used in training. We again localized the agent and patient regions using the leave-one-subject-out method described above, and averaged the "role-neutral" and "role-based" classification accuracies for each region.

In general, these results support our inference that these subregions are encoding role-specific information, rather than role-neutral semantic information. None of these regions are able to reliably identify the noun present, without respect to the role it occupies [anterior agent: $t_{(24)} = 0.50$, $P = 0.31$; posterior agent: $t_{(24)} = 0.24$, $P = 0.40$; patient: $t_{(24)} = -0.22$, $P = 0.41$].

The anterior and posterior agent regions trend toward being significantly better at identifying the agent than simply identifying the noun collapsing across roles [anterior: $t_{(24)} = 1.72$, $P = 0.097$; posterior: $t_{(24)} = 1.76$, $P = 0.09$]. Although the latter results are not statistically significant using a two-tailed test, we note that this is an exceptionally strong test. This is because a given noun, when present, will be present as the agent 50% of the time. The hypothesis that the agent region merely registers the presence of a given noun predicts that classification accuracy will be higher for general noun classification than for agent classification, and yet the pattern appears to be reversed. As expected, the patient region is significantly better at identifying the patient than identifying nouns in a role-neutral way [$t_{(24)} = 2.22$, $P = 0.036$].

Medial-lateral agent/patient topography of anterior lmSTC. Fig. 3*C* visualizes the representational content of anterior lmSTC moving along the medial-lateral axis. In performing these analyses, it was important to first localize the anterior lmSTC in a way that is unbiased with respect to its role preferences. To do this, we asked where in lmSTC a classifier could reliably tell, for a given noun (e.g., "man"), whether that concept was the agent or patient across trials. We conducted four such analyses, one for each noun, each time focusing only on trials in which the target noun was either the agent or patient. This analysis jointly localizes the set of subregions that encode the identity of the agent along with those that encode the identity of the patient. Because, for a given noun, being the agent is perfectly correlated with not being the patient and vice versa (in this analysis, although not in the experiment more broadly), both agent and patient regions should be identified by this analysis. Critically, this analysis reveals nothing about which type of information (agent/patient) is encoded in a particular subregion, making it a suitably unbiased localizer for present purposes.

For each of the four nouns, we again used an across-verb cross-validation procedure in which the classifier was trained on four of five verb contexts, and tested on the fifth, forcing it to generalize to data generated by new verb contexts. This was repeated for all four nouns, and the results were averaged. Given our goal of localizing a relatively large region, we used a liberal voxelwise threshold of $P < 0.05$ and cluster size of 250 voxels, which resulted in a significant cluster occupying a relatively large portion of the anterior lmSTC.

Within this ROI, we separately averaged the agent identification performance and the patient identification performance. We then computed the difference between the two performance levels at each Talairach X coordinate (this ranged from $X = -46$: -64), averaging over the anterior–posterior and superior–inferior axes. In other words, we examined performance levels in a series of slices, running along the medial–lateral axis. We then performed separate group-level t tests for agent identification accuracy, patient identification accuracy (Fig. 3*C*), and the difference between the two (Fig. 3*C*) at each X coordinate. These analyses show that medial regions of anterior lmSTC contain information about the identity of the agent, but not the patient. Lateral regions of anterior lmSTC contain information about the identity of patient, but not the agent. This selectively is reflected in the direct comparison of the two plotted in Fig. 3*C*.

"Deep" structure vs. linear order in lmSTC. To further confirm that the ROIs discovered by the searchlight analysis are indeed encoding the agent and patient of the propositions, we directly compared the performance of classification functions grouping sentences by their deep (agent/patient) structure to the performance of classifiers trained to group sentences by their linear order. When classifying based on deep structure, "the dog chased the man" and "the man was chased by the dog" were coded identically, as in our standard analyses. When classifying based on linear order, "the dog chased the man" and "the dog was chased by the man" were coded identically, given that the linear order of the words, and moreover the "surface subject" and "surface object" are the same across the sentences.

Using the leave-one-subject-out localization procedure described above, we found that classifiers trained to identify the underlying agent and patient performed significantly better in their respective subregions than classifiers trained to decode the surface subject [agent > surface subject: $t_{(24)} = 3.27$, $P = 0.003$] and surface object [patient > surface object: $t_{(24)} = 2.16$, $P = 0.046$], respectively. In fact, a subsequent search of lmSTC revealed no subregions that encoded information about the surface subject and surface object, as such.

Generalization between active and passive forms. The foregoing searchlight results demonstrate that there exist consistent patterns of activity for active and passive versions of the same proposition that classifiers can learn under supervision. We were also interested in whether classifiers trained to decode the agent or patient solely on one surface form can automatically generalize to the alternate form. We focused on the anterior agent, posterior agent, and patient ROIs identified by the across-verb searchlight procedure, and trained classifiers to make the same four-way decision (man? dog? cat? girl?) for each trial. In this case, the classifiers were trained only on active sentences, and tested on passive sentences, and then trained only on passive sentences and tested on active sentences. The results of these two procedures were then averaged and pooled across subjects.

We found that the neural representations in both the patient ROI and the posterior agent ROI automatically generalize across active and passive sentence forms [$t_{(24)} = 2.10$, $P = 0.023$; one-tailed: $t_{(24)} = 1.83$, $P = 0.039$], but that the anterior agent ROI patterns did not [$t_{(24)} = 1.10$, $P = 0.14$]. Although this difference may signal a functional difference between the two agent ROIs, it is important to note that this training procedure uses 50% of the data used by the searchlight analyses (288 trials vs. 144 trials), making it difficult to interpret this null result conclusively. To further evaluate whether this null effect in generalizing across active and passive voice for the agent region is due to limited power, we reran the principal agent identification analysis, which subsumes active and passive sentences under the same function, using a random 50% of each subject's data. This matches the amount of data available when generalizing automatically across the active and passive voice. We found that agent classification subsuming active/passive is significantly better than agent classification when generalizing from active to passive (and vice versa), even when the number of observations used to train the classifier are matched [$t_{(24)} = 2.3$, $P = 0.03$]. Although this effect is not large, it suggests that the agent region's failure to automatically generalize across voices may reflect a real difference in the representation of the agent when communicated through the active and passive voices. That is, it suggests that these regions may not simply be sensitive to "Who did it?" and "To whom was it done?" but also to the way in which the value of these variables is derived from the sentence's syntax.

Syntactically, the patient is the verb's internal argument (that is, internal to the verb phrase) for both active and passive sentences. Accordingly, we find that representations of the patient generalize automatically across voices. However, the agent is an external argument when expressed in the active voice, but part of the nonmandatory "by" phrase (e.g., "was chased by the dog") when expressed in the passive voice. Thus, although the patient is an argument in both voices, the syntactic status of the agent differs across voices. This different mapping from syntax to semantic role may influence the classifier's ability to automatically generalize across voices for the agent region, even if the region can learn a function that groups the agent in active and passive sentences together, as we see in our principal analysis. Whether these regions carry information about the structure of the sentence is an important topic for future research.

Classification performance by verb context and noun. The foregoing searchlight analyses averaged classification performance across the four nouns to-be-identified and generalization ability to patterns generated by the five verb contexts. It thus remains possible that the results owe to particularly strong performance on a subset of nouns and/or verb contexts, with no information about the others. Whether the results are driven by a subset of nouns and verb contexts is of considerable interest to the interpretation of the results: Do these regions house domain-general mechanisms for encoding structured semantic content, used across nouns and verb contexts? Or do the regions specialize in particular semantic content, either in the semantic content of the nouns they represent or the verb contexts in which they appear? We were therefore interested in whether the classifiers performed consistently when separately analyzed for each noun and each verb context.

To determine whether the neural representations generalize to all verbs used, we performed a searchlight analysis in which local classifiers were (*i*) trained to make the same four-way agent/patient identification decisions described above using data generated by the target verb (20% of the data) and asked to generalize to the remaining four verbs (80% of the data), and (*ii*) trained using the remaining four verbs (80% of the data) and asked to generalize to the target verb (20% of the data). The results of *i* and *ii* were then averaged to provide a measure of how well the patterns of activity corresponding to noun/role combinations in that verb context generalize to the other four verb contexts. This analysis produced five separate search maps of lmSTC for agent/patient identification in generalizing to each of the five verbs. We iteratively localized clusters containing information for four of five verbs using these search maps and a liberal threshold ($P < 0.05$ voxelwise, 50 mm^3) and asked whether the average classification accuracy in generalizing to the fifth verb in the identified region was significantly greater than chance. Table S4 shows the results of these analyses for the three ROIs, by generalization to each verb.

We then performed a similar analysis examining classification performance, broken down by the nouns occupying these roles. Here, we performed separate searchlight analyses for each of the six possible pairwise noun discriminations (e.g., discriminating "man as agent" vs. "girl as agent"), for both the agent and patient roles. The ROIs were again iteratively localized using the results of five of six pairwise classifications and a liberal threshold ($P < 0.05$ voxelwise,

50 mm^3), and the searchlight results of sixth pair were averaged across the resulting ROI. The results are presented for each pairwise comparison in Table S4.

Experiment 3: Replication of Experiment 2

Experiment 3 was conducted using the same facilities, equipment, and parameters as experiments 1 and 2 with changes identified below.

Data Acquisition, Subjects, and Preprocessing. Each functional echo-planar imaging volume consisted of 58 slices parallel to the anterior commissure (field of view, 192 mm; repetition time, 3,500 ms; echo time, 28 ms; flip angle, 90°). We used parallel imaging (iPAT 2) to obtain whole-brain coverage with 2 × 2 × 2-mm voxels. We analyzed data from 41 participants aged 18–35. The data were not smoothed before classification analyses.

Stimuli and Experimental Procedure. As in experiment 2, sentences were constructed from a menu of nouns and verbs to form every possible aRb proposition with two distinct nouns. We used six monosyllabic English nouns that refer to animals: "moose," "cow," "hog," "crow," "goose," and "hawk." We used eight transitive verbs: "chased," "approached," "passed," "attacked," "surprised," "frightened," "noticed," and "detected." For both motion verbs (e.g., "chased" and "attacked") and psychological verbs (e.g., "surprised" and "noticed"), we classified the entity that caused the event as the agent and the other entity as the patient. That is, we treated the active-voice objects of "surprised" and "frightened" and the active-voice subjects of "noticed" and "detected" as the patients. Each proposition was presented only once. Whether a proposition was presented in the active or passive voice was randomly determined, and 50% of the propositions were presented in each voice. Each sentence was presented visually for 3.5 s, followed by 7 s of fixation. After 7 s, participants could be asked a question about who did what to whom in the sentence they had just read.

Searchlight Analyses. The search region of interest for experiment 3 was formed by dilating the group-level anterior agent and patient ROIs discovered in experiment 2 by 8 mm. All classification analyses were conducted in the subject's native space. As in experiments 1 and 2, all used a linear classifier with a shrinkage estimate of the covariance matrix. Local voxel neighborhoods were defined using a 2-mm (one-voxel) radius, entailing that nonedge neighborhoods contained 27 voxels.

For all classification analyses, searchlights iteratively selected data from local voxel neighborhoods to make six-way decisions regarding the noun occupying the agent or patient role on a given trial. Individual-level accuracy maps were smoothed with a 2-mm FWHM kernel, warped to Talairach space. The group of subjects' maps was submitted to a directional one-sample t test against 0.16667 (chance accuracy) to determine whether any regions reliably encoded information about the identity of the agent or patient across subjects. We used a liberal voxelwise threshold of $P < 0.05$ coupled with a Monte Carlo simulation to determine the probability of obtaining significant clusters given data containing only noise (clusterwise corrected threshold of $P < 0.05$). As in experiments 1 and 2, we estimated the smoothness of the data by conducting the same classification and aggregation procedures with randomly permuted labels.

Replication Results. As in experiment 2, we find a medial region of lmSTC that carries information about the identity of the agent ($k = 37$, $P < 0.05$, clusterwise corrected), and a lateral region that carries information about the identity of the patient ($k = 107$, $P < 0.001$, clusterwise corrected) (Fig. S5).

Also as in experiment 2, we next localized these regions in $N - 1$ subjects, leaving each subject's data out, and tested the mean classification accuracy across these agent and patient regions in the withheld subject. As in experiment 2, we find a significant role-by-region interaction [$F_{(1,40)} = 8.7$, $P < 0.005$; permutation test: $P = 0.0045$]. The patient region is again the more selective region: Patient classification is significantly better than agent classification within the independently localized patient regions [$t_{(40)} = 2.48$, $P = 0.017$; permutation test: $P = 0.017$]. Agent classification is significantly better than chance within the agent regions [$t_{(40)} = 1.91$, $P = 0.03$; permutation test: $P = 0.038$], whereas patient classification is not [$t_{(40)} = -0.13$, $P = 0.45$; permutation test: $P = 0.5$]. As in experiment 2, agent classification is not significantly better than patient classification within the agent region [$t_{(40)} = 1.26$, $P = 0.21$; permutation test: $P = 0.22$].

These results thus replicate those of experiment 2. We again find that a lateral region of lmSTC encodes the identity of the patient, whereas a medial region of lmSTC encodes the identity of the agent. We once again find that the patient region is selective for patient information. As before, the evidence for selectivity in the agent region is suggestive but somewhat weaker.

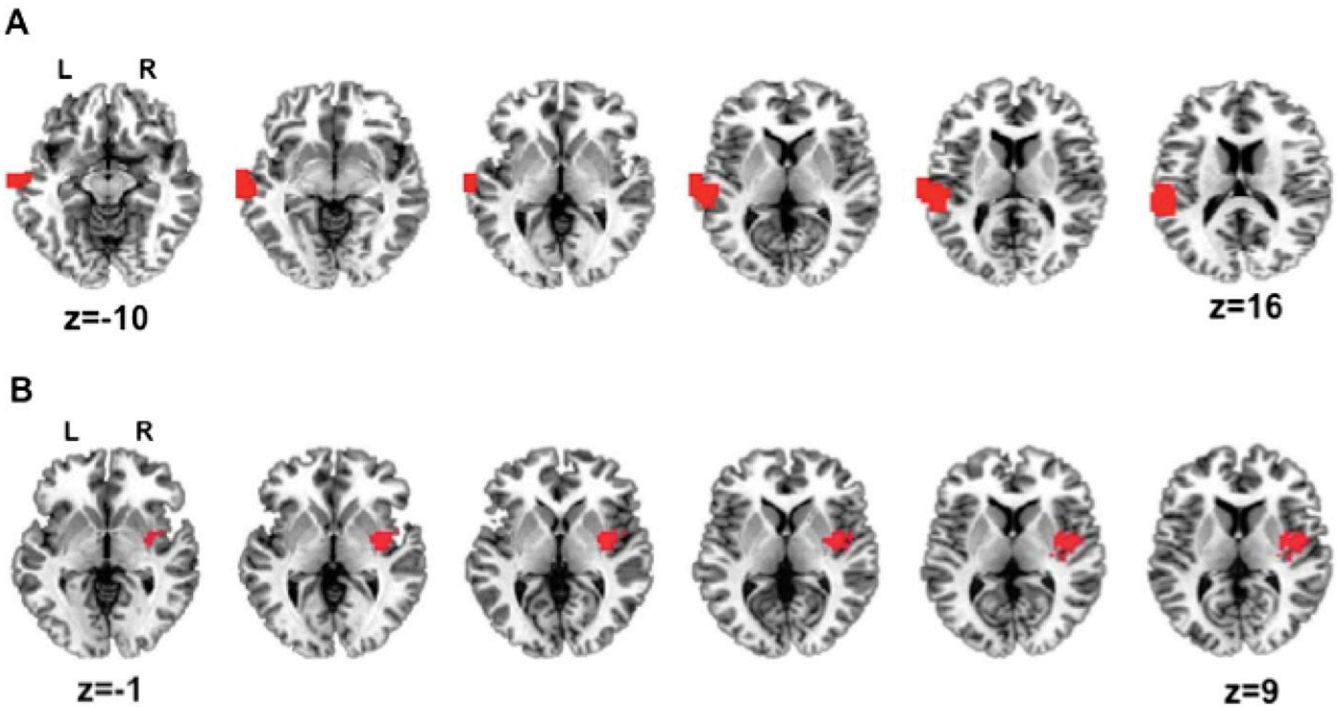

Fig. S1. Experiment 1 whole-brain searchlight analysis results. (*A*) Left mid-superior temporal cortex (lmSTC) and (*B*) right insula/extreme capsule region found to distinguish mirror image propositions (e.g., "the truck hit the ball"/"the ball the truck").

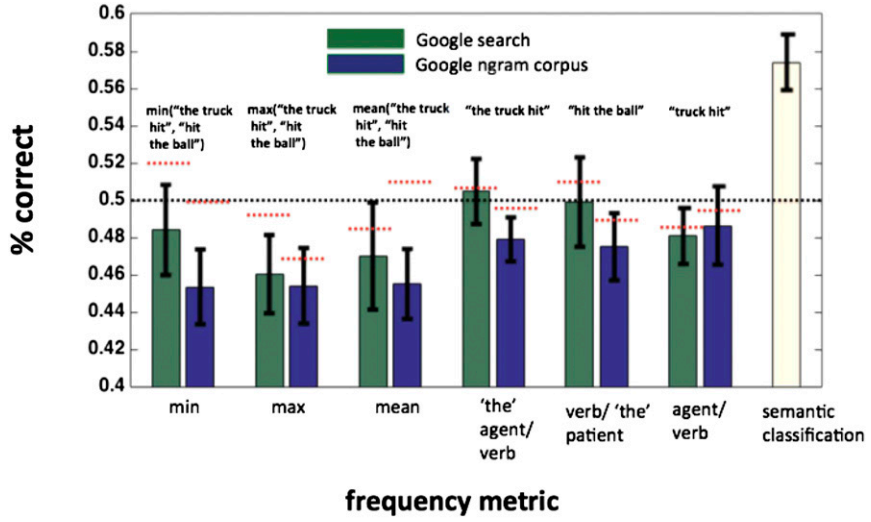

Fig. S2. As a control, we attempted to use lmSTC patterns of activity to predict frequency statistics for noun/verb and verb/noun combinations (e.g., "the truck hit," or "hit the ball"). The patterns of activity did not reliably carry information about any of the frequency statistics tested. Red horizontal bars indicate the mean results of a randomization procedure that scrambled the frequency statistics and performed the same analysis 1,000 times for each subject. The text above each bar pair provides an example of the part of an example proposition, "the truck hit the ball," described by that frequency statistic.

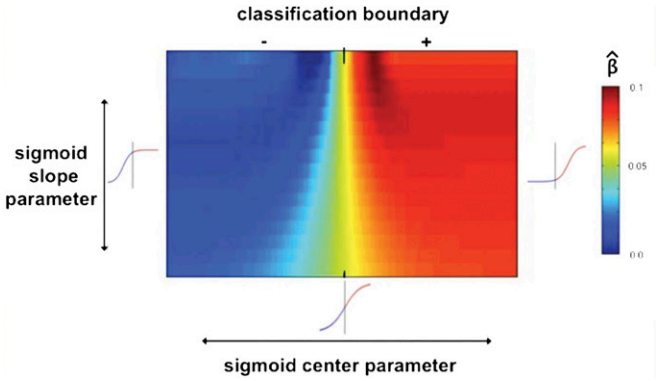

Fig. S3. Heat map visualizing variation in the prediction of amygdala activity as a function of different sigmoidal transformations of the signed distance of an observation from the lmSTC classification boundary. Colors correspond to different β values for this regression averaged across 16 cross-validation iterations. The *x* axis represents different values of the sigmoid's "center." The *y* axis represents different slopes, with steeper slopes located at the top of the graph. Together, these two parameters yield different shapes, three of which are shown at the appropriate locations in the space for reference. Vertical bars through the reference sigmoids correspond to the location of the classification boundary, as does the bar at the top along the *x* axis. The "+" signifies the positive side of the classification hyperplane (the "grandfather kicked the baby" side), whereas the "−" signifies that negative side of the hyperplane. We see better prediction when the sigmoid is centered at or to the right of the hyperplane. The best prediction is obtained by a function with a steep slope, centered slightly to the right of the hyperplane. (See Fig. 1*B*.) This demonstrates that patterns corresponding to "the grandfather kicked the baby" are more likely to elicit an amygdala response than their mirror image, as explained in the main text. However, the rightward shift of the best sigmoid, past the classification boundary, also suggests that the likelihood of no amygdala response, given the "grandfather kicked the baby" pattern in lmSTC was greater than the likelihood of an amygdala response, given the "baby kicked the grandfather" pattern. More generally, this demonstrates that the point of indifference for classification need not coincide with the point of indifference for predicting a response elsewhere in the brain. This approach may be useful for empirically testing subtly different quantitative functions relating information in one brain area to information or activation in other.

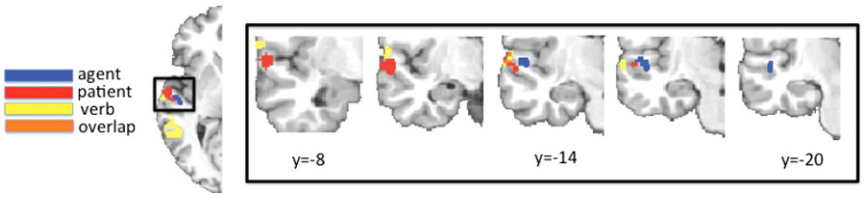

Fig. S4. Searchlight results for experiment 2 for the three classification problems. Coronal slices show the topography of the neighboring anterior verb, agent, and patient subregions.

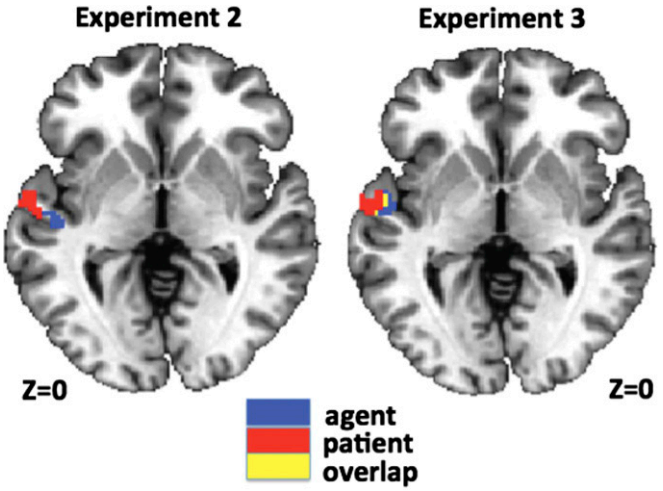

Fig. S5. Agent and patient searchlight results within lmSTC for experiments 2 and 3. In both experiments, we identify a medial region that encodes information about the agent (blue) as well as an adjacent lateral region that encodes information about the patient (red). For both experiments, images show clusters surviving a threshold of $P < 0.05$, corrected. For experiment 3, analysis was restricted to a region created by dilating the regions identified in experiment 2 by 8 mm.

Table S1. Active and passive versions of the sentences used in experiment 1

Pair	Mirror image sentence 1	Mirror image sentence 2	Use in experiment 1
Active voice			
1	The grandfather kicked the baby	The baby kicked the grandfather	Connectivity
2	The mother struck the boy	The boy struck the mother	Connectivity
3	The father pulled the child	The child pulled the father	ROI localization
4	The grandmother touched the girl	The girl touched the grandmother	ROI localization
5	The truck hit the ball	The ball hit the truck	ROI localization
6	The door smacked the branch	The branch smacked the door	ROI localization
Passive voice			
1	The baby was kicked by the grandfather	The grandfather was kicked by the baby	Connectivity
2	The boy was struck by the mother	The mother was struck by the boy	Connectivity
3	The child was pulled by the father	The father was pulled by the child	ROI localization
4	The girl touched the grandmother	The grandmother touched the girl	ROI localization
5	The ball was hit by the truck	The truck was hit by the ball	ROI localization
6	The branch was smacked by the door	The door was smacked by the branch	ROI localization

Table S2. Post hoc analysis of lmSTC and right posterior insula ROIs discovered by the whole-brain searchlight analysis

Pair	Mirror image pair	Mean accuracy
lmSTC		
1	STRUCK (mother, boy)	0.566
2	KICKED (grandfather, baby)	0.602
3	TOUCHED (grandmother, girl)	0.563
4	PULLED (father, child)	0.565
5	HIT (truck, ball)	0.556
6	SMACKED (door, branch)	0.597
Right insula		
1	STRUCK (mother, boy)	0.562
2	KICKED (grandfather, baby)	0.557
3	TOUCHED (grandmother, girl)	0.583
4	PULLED (father, child)	0.546
5	HIT (truck, ball)	0.565
6	SMACKED (door, branch)	0.562

Classification performance is broken down by mirror image proposition pair.

Table S3. Searchlight results for experiment 2 listing significant clusters within lmSTC

Identification problem	Center coordinates (Talairach)	No. voxels	Smoothness of noise map; *XYZ*, mm	Corrected *P*(data|noise)
Agent	−46, −18, 1	67	4.27, 5.72, 4.93	$P < 0.01$
	−57, −37, 7	61		$P < 0.02$
Patient	−57, −10, 2	181	5.13, 5.6, 5.47	$P < 0.0001$
Verb	−61, −15, 2	60	4.59, 6.17, 5.08	$P < 0.025$
	−55, −49, 5	433		$P < 0.0001$

All searchlight analyses used a voxelwise threshold of $P < 0.005$, and a corrected threshold of $P < 0.05$ to identify significant clusters.

Table S4. Post hoc analyses of agent and patient identification performance within their corresponding ROIs when generalizing to each verb context and new pairwise noun discriminations

Content	Patient in patient ROI	Agent in anterior agent ROI	Agent in posterior agent ROI
Generalization by verb context			
Verb			
"Chase"	2.12, $P = 0.022$	2.04, $P = 0.026$	1.59, $P = 0.063$
"Block"	1.67, $P = 0.054$	2.37, $P = 0.013$	2.16, $P = 0.021$
"Bump"	3.03, $P = 0.003$	1.78, $P = 0.044$	2.37, $P = 0.013$
"Approach"	2.15, $P = 0.021$	2.61, $P = 0.008$	3.19, $P = 0.002$*
"Scratch"	4.31, $P = 0.0001$*	2.77, $P = 0.005$	1.87, $P = 0.037$
Generalization by noun pair			
Noun pair discrimination			
Man/Girl	1.85, $P = 0.039$	2.68, $P = 0.0065$	3.21, $P = 0.0019$*
Man/Dog	1.23, $P = 0.115$	1.47, $P = 0.077$	2.10, $P = 0.023$
Man/Cat	2.90, $P = 0.004$	2.22, $P = 0.018$	2.02, $P = 0.027$
Girl/Dog	1.20, $P = 0.121$	1.08, $P = 0.145$	2.62, $P = 0.0075$
Girl/Cat	3.53, $P = 0.0009$*	3.2, $P = 0.0019$*	1.56, $P = 0.066$
Dog/Cat	1.71, $P = 0.05$	1.20, $P = 0.121$	1.06, $P = 0.15$

Values are *t* statistics, with one-tailed *P* values against chance. Our main goal is to qualitatively describe the pattern of results. However, asterisks (*) mark accuracies that withstand Bonferroni correction for multiple comparisons.