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

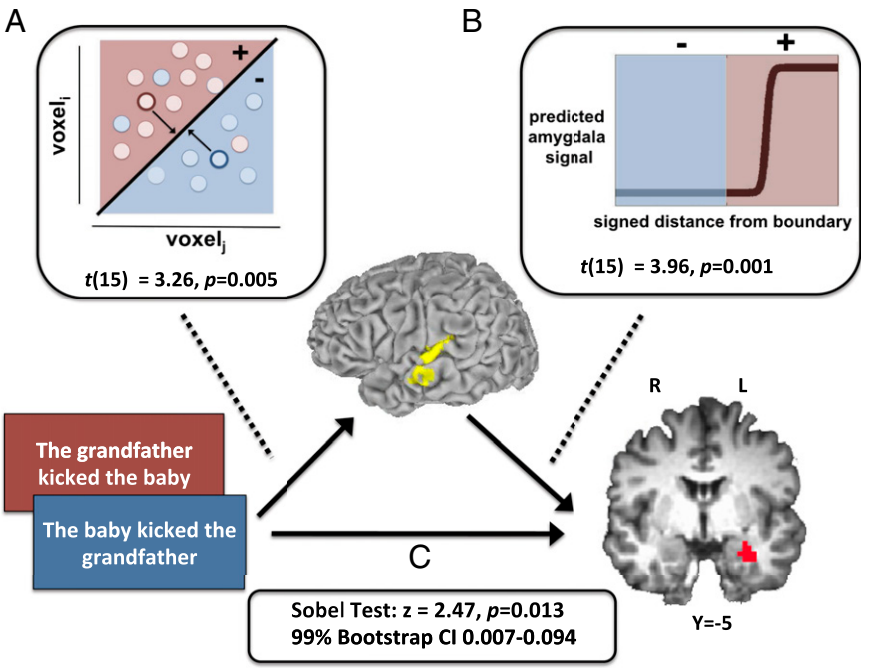

Fig. 1. Model of information flow from stimulus to lmSTC to amygdala in experiment 1. (*A*) A pattern classifier determines which of two propositions was presented using activity in lmSTC. Distance from the classification boundary indicates the extent to which a learned pattern was instantiated. The 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%).

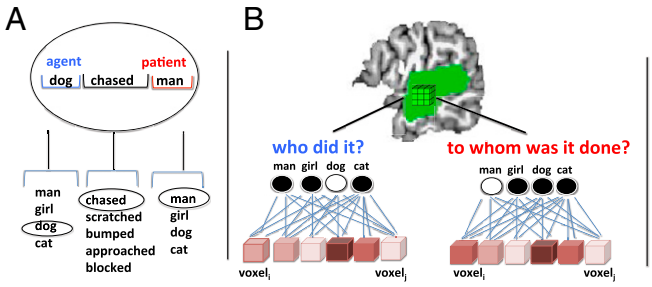

Fig. 2. Experiment 2 design. (*A*) Subjects read sentences constructed from a menu of five verbs and four nouns, with one noun in the agent role and another in the patient role. (*B*) For each trial, separate pattern classifiers attempted to identify the agent and the patient based on activity within 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

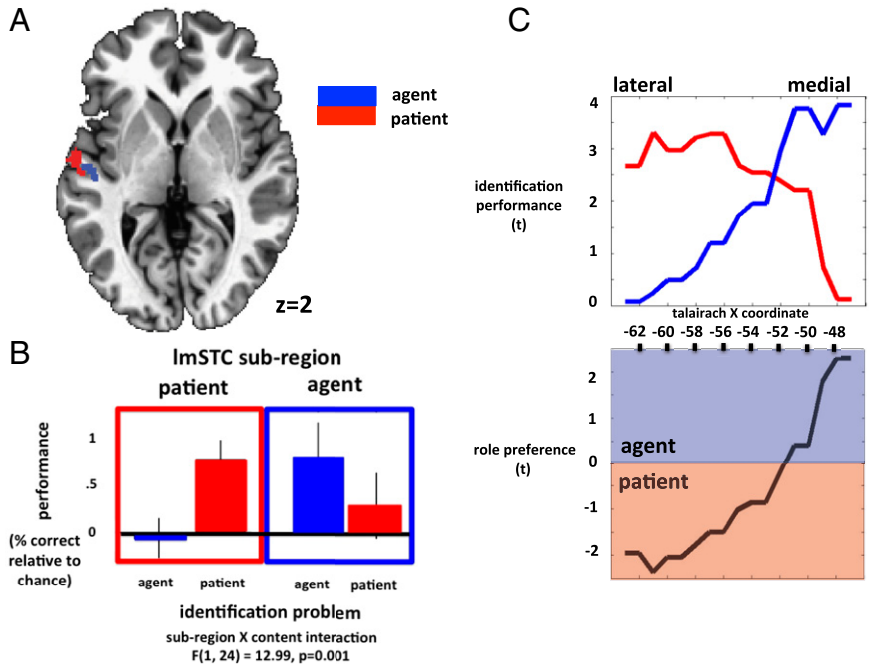

Fig. 3. (*A*) Searchlight analyses identified adjacent, but nonoverlapping subregions of anterior lmSTC that reliably encoded information about agent identity (medial, blue) and patient identity (lateral, red). (*B*) Post hoc analyses find that these adjacent regions differ significantly in the information 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.