Cerebral Cortex, 2020;00: 1–18

doi: 10.1093/cercor/bhaa001
Original Article

ORIGINAL ARTICLE

Two Ways to Build a Thought: Distinct Forms of Compositional Semantic Representation across Brain Regions

Steven M. Frankland[1] and Joshua D. Greene[2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, and [2]Department of Psychology, Center for Brain Science, Harvard University, Cambridge, MA 02138.

Address correspondence to: Steven M. Frankland, Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA. Email: steven.frankland@princeton.edu

Abstract

To understand a simple sentence such as "the woman chased the dog", the human mind must dynamically organize the relevant concepts to represent who did what to whom. This structured recombination of concepts (woman, dog, chased) enables the representation of novel events, and is thus a central feature of intelligence. Here, we use functional magnetic resonance (fMRI) and encoding models to delineate the contributions of three brain regions to the representation of relational combinations. We identify a region of anterior-medial prefrontal cortex (amPFC) that shares representations of noun-verb conjunctions across sentences: for example, a combination of "woman" and "chased" to encode woman-as-chaser, distinct from woman-as-chasee. This PFC region differs from the left-mid superior temporal cortex (lmSTC) and hippocampus, two regions previously implicated in representing relations. lmSTC represents broad role combinations that are shared across verbs (e.g., woman-as-agent), rather than narrow roles, limited to specific actions (woman-as-chaser). By contrast, a hippocampal sub-region represents events sharing narrow conjunctions as dissimilar. The success of the hippocampal conjunctive encoding model is anti-correlated with generalization performance in amPFC on a trial-by-trial basis, consistent with a pattern separation mechanism. Thus, these three regions appear to play distinct, but complementary, roles in encoding compositional event structure.

Key words: compositionality, encoding models, fMRI, meaning, memory

Introduction

To understand the meaning of a novel sentence, our brains rely on the principle of compositionality: a sentence's meaning is a function of 1) the meanings of its parts and 2) the way in which those parts are combined (see Frege and Patzig 2003; Montague 1970; Fodor and Pylyshyn 1988). Understanding even simple sentences, such as "the woman chased the dog" requires not only retrieving knowledge about dogs, women, and chasing, but also combining these representational elements in a way that reflects the relational structure of the particular event described: did the dog chase the woman, or the other way around?

Over the last two decades, many studies have used functional neuroimaging to examine the brain's encoding strategies for representing the meanings of a sentence's parts. This includes work on reusable object knowledge (e.g., Chao et al. 1999; Thompson-Schill 2003; Mitchell et al. 2008; Fairhall and Caramazza 2013) and action/event knowledge (Kemmerer et al. 2008; Bedny et al. 2008; Peelen et al. 2012; Huth et al. 2012; Elli et al. 2019), as well as broader attempts to map semantic representations across cortex (Mitchell et al. 2008; Huth et al. 2016). However, far less is known about how the brain combines word-level meanings to flexibly encode the meaning of a particular sentence, even though this type of combinatorial process

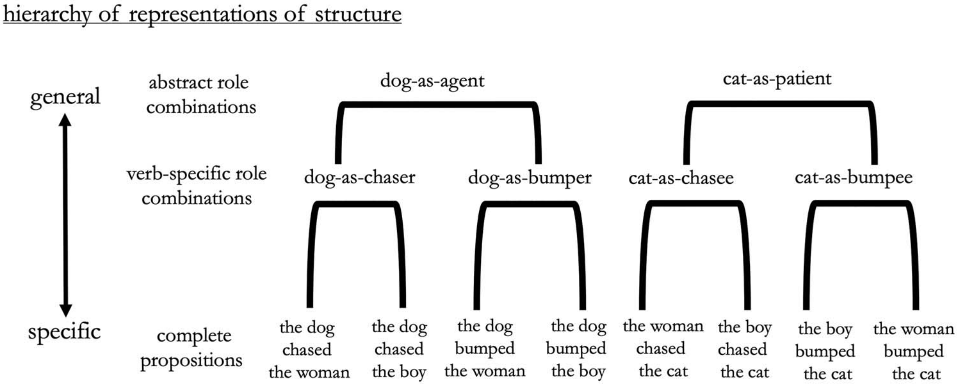

Figure 1. Understanding simple descriptions of events requires encoding the relations between the event's participants (who did what to whom). Consider the proposition "the dog chased the cat." This proposition can employ structured representations from at least two distinct levels of a hierarchy. First, conjunctions of specific noun-verb combinations ("the dog chased," "the cat was chased") can be reused across propositions involving the same verb and the same noun in the same relationship ("the dog chased the man" and "the dog chased the cat"). At a higher level of abstraction, however, there are semantic role representations that can generalize across verbs. For example, dog as the agent (the entity that is causally responsible for affecting another entity) can be a thing that bumps something or chases something ("the dog chased the cat" "the dog bumped the boy"). Although narrow role combinations (bindings of nouns to specific verbs) are invariant to the remaining arguments of the relation ("the dog chased [something]"), broader role combinations are invariant to both the remaining argument, and the particular verb ("the dog did [something]"). Thus, they exist at a higher level of abstraction.

is central to high-level cognition (Fodor and Pylyshyn 1988; Smolensky 1990; Plate 1995; Pinker 1997; Hummel and Holyoak 2003; Doumas et al. 2008). Although a considerable body of work has identified perisylvian regions engaged in complex semantic and syntactic processing (Mazoyer et al. 1993; Vandenberghe et al. 2002; Humphries et al. 2006; Fedorenko et al. 2011; Pallier et al. 2011), it remains unclear how the time-varying relational representations necessary to encode sentence meanings (such as who did what to whom) are encoded in patterns of neural activity. Here, we focus on a particular aspect of this question, identifying and characterizing brain regions that are sensitive to the roles particular entities play in an event.

We consider two distinct representational strategies for encoding who did what to whom, differing in their level of abstraction (Fig. 1). First, structured events can be encoded by assigning noun-meanings to broad semantic roles that are reused across verbs. For example, to encode the woman chased the dog, the meaning of "woman" may be assigned to the agent role (the entity that does something), and the meaning of "dog" may be assigned to the patient role (the entity that has something done to it). We call these "broad" roles because the representation is invariant across a broad class of events. A woman, qua agent, may do many things: climb, scratch, jump etc. Such abstract role representations are well suited to mapping event structure onto syntactic structure, as agents tend to be subjects, and patients tend to be objects (Van Valin Jr and Van Valin 2005; Levin and Hovav 2005). Thus, these broad semantic roles are thought to play an important role in language acquisition and use.

Broad roles, however, abstract away from information about how particular noun and verb meanings interact in an event. For example, to generate a mental image of "a chasing woman," the act of chasing and the chasing agent must be integrated into a coherent representation, such that the woman looks different when she is chasing as opposed to, say, climbing. To maintain this information, the system might use narrower semantic roles, specific to a particular event-type. For example, "the woman chased the dog" may be represented as a composition of two event-specific conjunctions, such as woman-as-chaser and dog-as-chasee (Selfridge 1958). Here, we focus primarily on the simple distinction between verb-specific (which we call "narrow") and verb-invariant (which we call "broad") roles. Note, however, that we do not intend our characterization of these roles to be exhaustive, but rather a first pass at delineating those neural systems that reflect a trade-off between abstraction and specificity in role representation. Within the linguistics literature, broad roles have themselves been suggested to exist at various levels of abstraction, ranging from classical semantic roles (Fillmore 1967), such as agent and patient, to more abstract macro-roles such as "actor" and "undergoer" (Van Valin, Jr and Van Valin 2005; see also Dowty 1991 on "proto-roles"). There are likely more than two ways to build a thought.

These two different ways of building a thought—using broad vs. narrow semantic role combinations—thus trade-off abstraction (generality) for specificity (information). Thus, these representations may serve different functions, not only in language learning (Pinker 1989; Tomasello 1992; Goldberg 1995; Gertner et al. 2006), but in cognition more generally. Moreover, broad and narrow representations are associated with different cognitive architectures: neural networks trained with backpropagation typically learn narrow feature conjunctions reflecting the statistical structure of the training domain, while classical architectures often impose prior structure to favor abstract variables typical of computer programs (Pinker 1997; Marcus 2001). To better understand how the human brain represents the relations necessary to understand sentence meaning, we use functional magnetic resonance (fMRI) and voxelwise encoding models to ask which strategies the brain uses—broad roles vs. narrow roles—and whether different regions employ different strategies.

We begin with a whole-brain search for regions whose activity generalizes to new sentences containing familiar parts, consistent with the reuse of representations across sentences. This analysis identifies a region of anterior-medial prefrontal cortex (amPFC) that reflects event structure, differentiating reversed pairs containing the same parts ("the woman chased the dog" vs. "the dog chased the woman"). We then use a set of more specific encoding models to characterize the representational profile of this region: Does it reuse semantic components in which the meanings of nouns are bound to broad roles, narrow roles, or both?

We also apply such encoding models to two a priori ROIs, previously implicated in representing relations within an event, the lmSTC (Wu et al. 2007; Frankland and Greene 2015) and the hippocampus (Cohen and Eichenbaum 1993; Davachi 2006; Libby et al. 2014; Duff and Brown-Schmidt 2012). lmSTC has been found to carry information about who did what to whom in sentences (Frankland and Greene 2015) and nearby regions carry information about who did what to whom in videos (Wang et al. 2016). Moreover, damage to lmSTC produces deficits in tasks requiring thematic role assignment (Wu et al. 2007). The hippocampus, by contrast, is thought to incorporate contextual information to rapidly and flexibly bind separate elements of an event (Cohen and Eichenbaum 1993; Eichenbaum 1999). Evidence suggests that hippocampus is particularly integral to encoding relations between these elements, rather than elements themselves (Cohen and Eichenbaum 1993; Davachi 2006; Ranganath and D'Esposito 2001). Given that sentence comprehension, too, requires flexibly encoding relations between distinct elements, researchers have suggested that the hippocampus may be well-suited to contribute to dynamic aspects of language comprehension (Duff and Brown-Schmidt 2012; Blank et al. 2016; Piai et al. 2016).

To foreshadow our primary results, we find that an anterior-medial region of prefrontal cortex represents "narrow," reusable subparts of sentence meanings (e.g., woman-as-chaser as part of "woman chases dog"). The narrow conjunctive encoding model's success in this prefrontal region is anticorrelated with its success in the hippocampus, which contains a subregion that appears to separate representations of sentences sharing these noun-verb conjunctions. In contrast, lmSTC reuses broad noun-role combinations, shared across verbs (e.g., woman-as-agent), tracking the abstract structure of the event that is common across the class of verbs studied. Critically, both broad and narrow representational forms generalize across sentences, enabling the representation of unfamiliar, structured events that involve familiar pieces. Critically, these complementary strategies exploit representations at different levels of abstraction.

Although our primary focus is on understanding the representations that enable the encoding and interpretation of who did what to whom in novel events, our stimulus set dissociates the semantic and phonological similarity of the nouns and the semantic and syntactic roles of the verbs (see Fig. 2), enabling us to probe the representational content of these regions in post hoc analyses.

Materials and Methods

Stimuli and Procedure

Sentences were constructed from a menu of 6 nouns and 8 transitive verbs (see Fig. 2B), creating every possible subject-verb-object combination, excluding propositions in which the same noun occupied both roles (e.g., "the goose approached the goose"). The particular set of nouns and verbs were constructed so that we could, in exploratory analyses, dissociate semantic from phonological codes and semantic from syntactic structure, in regions of interest. These aspects of the stimuli (see Figs 2D and 4A) are described in detail below ("Similarity & Structure: ROI analyses.").

These sentences were intended to be unfamiliar to subjects. None of the active sentences were found in Google's 5gram corpus (https://catalog. ldc.upenn.edu/LDC2009T25), and no instances of the active or passive sentences were returned via Google search at the time of the experiment, suggesting that subjects were unlikely to have encountered these particular combinations, even though the individual nouns and verbs were familiar.

Over the course of the experiment, subjects thus read 240 unique propositions while undergoing fMRI, each presented once. The 240 sentences were evenly and randomly distributed over six scan runs. Whether a proposition was presented in the active or passive voice was randomly determined for each subject. Each run contained the presentation of 40 sentences. Each sentence was visually presented for 3.5 s (1 TR) followed by 7 s of fixation (2 TRs). On one third of the trials, randomly chosen, a comprehension question followed the fixation period. These questions were of the form "Did the hawk approach something?" or "Was the moose approached by something?", and thus only required encoding the event participants in terms of the abstract structural roles they occupied. In total, 50% had affirmative correct answers.

Data Collection and Subjects

The experiment was conducted using a 3.0 T Siemens Magnetom Tim Trio scanner with a 32-channel head coil at the Harvard Brain Sciences Center in Cambridge, MA. A high-resolution structural scan (1 mm^3 isotropic voxel MPRAGE) was collected prior to functional data acquisition. Each functional EPI volume consisted of 58 slices parallel to the anterior commissure (FOV = 192 mm, TR = 3500 ms, TE = 28 ms, Flip Angle = 90°). We used parallel imaging (iPAT 2) to obtain whole-brain coverage with 2 × 2 × 2 mm voxels. Stimuli were presented using Psychtoolbox software (http://www.psychtoolbox.org) for Matlab (http://www.mathworks.com).

In total, 55 members (24 male, 31 female, aged 18–32 (M = 22.9) of the Cambridge, MA community participated for payment. All subjects were native English speakers, self-reported right handed, had normal or corrected-to-normal vision, and gave written informed consent in accordance with Harvard University's institutional review board. Subjects had a mean accuracy of 84.9% (SD = 0.09) for the comprehension task (chance performance =50%). Those subjects (N = 5) who performed below 70% on the comprehension task were excluded from data analysis. Data from two additional subjects were not analyzed due to excessive movement. This left 48 subjects remaining for analyses (a subset of the present data was used for a distinct analysis reported in the supporting information of Frankland and Greene (2015). Those supplemental analysis replicated the analysis and findings reported in Experiment 2 of that paper. None of the results herein were previously reported.).

Data Analysis

Preprocessing

Image preprocessing was performed using AFNI functions (Cox 1996) and custom scripts, implemented in Matlab (http://www.mathworks.com). Each subject's EPI images were spatially registered to the first volume of the first experimental run. Motion parameters, global signal across the brain, and first, second, and third order temporal trends were removed from each voxel's time course. Data were then smoothed with a Gaussian kernel at 2 mm FWHM. Following Mumford et al. (2012), we modeled each trial (here, the sentence presentation) using a generic regressor,

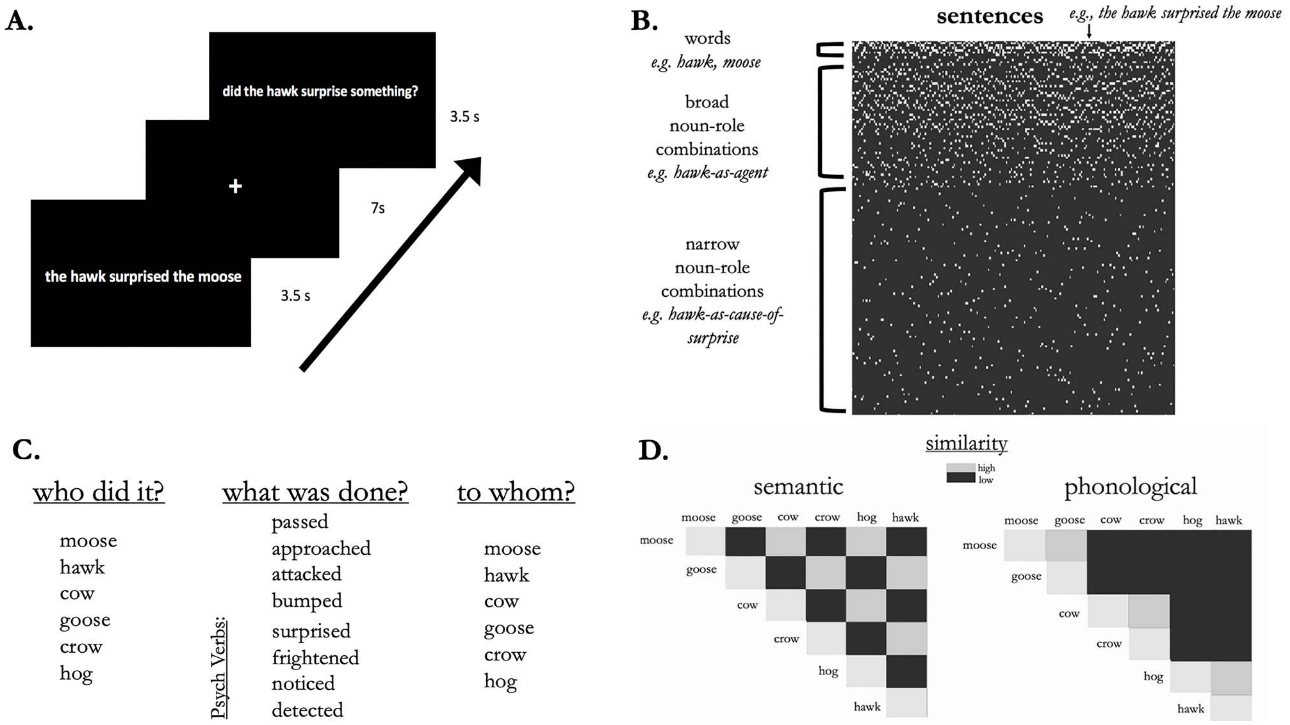

Figure 2. (A) While undergoing fMRI, subjects read simple sentences describing events, and were asked to remember the sentence's meaning for a short delay period. (B) We modeled the BOLD signal during sentence presentation as a linear combination of these reusable sentence components (nouns, verbs, specific noun-verb combinations, and noun-role combinations) and asked where in the brain the model could predict neural activity to unfamiliar sentences sharing these components. In the model, each particular sentence (column) is coded as a binary vector reflecting the presence or absence of recurring sentence components (rows). These variables ranged from the presence of words in the sentence, without respect to the role that word played (e.g., that the noun 'moose' was included, ignoring relational structure) to broad noun-role combinations (moose-as-agent shared across verbs), to narrow noun-role combinations, combining nouns and specific verbs in a specific relation (e.g., "moose-as-surpriser something"). (C) Sentences were constructed from a menu of 6 nouns and 8 verbs. (D) These nouns were selected because they can be described with dissociable semantic and phonological similarities spaces. This enables us to study the encoding schemes employed in our ROIs.

convolved with a canonical hemodynamic response function, provided in SPM. All other trials in the run, including comprehension questions, were included in the regression as covariates of no interest. This produces one beta value for each sentence at each voxel, reflecting the BOLD response to that sentence (trial). These trial-by-trial, sentence-specific beta estimates were used as data for all analyses.

General Encoding-Model Analysis Procedures

Encoding models were trained to predict BOLD signal at each voxel as a weighted, linear combination of sentence descriptors (see Fig. 2B). The parameters were fit to data using a subset of sentences, and then used to predict neural activity to sentences withheld from the model training. We used k-fold cross-validation, with scan runs treated as folds. We describe the various sentence models below, but here focus on those analysis procedures shared across models.

For each cross-validation iteration, the model was trained on data from 5 of 6 scan runs and tested on data from the held-out run. Thus, each training iteration used 200 of the 240 unique sentences to fit model parameters, and its predictions were evaluated on the remaining 40 sentences. The β parameters of the voxelwise encoding model were fit separately for each subject, each voxel, and each cross-validation iteration as least squares estimates in a multiple regression. Given that the number of model parameters was always less than the number of observations, an additional regularization penalty was not necessary.

We evaluated the model's performance using the following procedure. For each cross-validation iteration, we used the learned parameters to generate a prediction for each voxel for each of the 40 held-out sentences. For a given voxel and cross-validation iteration, the predicted data and observed test data are both 1 × 40 vectors (predictions and observations for that 1 voxel × 40 held out trials). Using these, we construct a 40 × 40 matrix populated by the squared differences (errors) between these 40 predictions and 40 observations (see Supplementary Fig. 1). The on-diagonal elements in this matrix contain the correct mappings between predicted and observed data. The off-diagonal elements contain the incorrect mappings. To evaluate the model for a particular iteration and voxel, we z-score over the entire error matrix of squared differences for that iteration, and ask whether the average of the on-diagonal elements (correct mappings) is lower than that of the off-diagonal elements (incorrect mappings). For example, the difference between the predicted BOLD signal for the sentence "the cow approached the crow" should be more similar to the observed BOLD signal for that sentence, as compared to the observed signal for other sentences, for example, the sentence "the hawk attacked the cow." These difference scores were then averaged across the 6 cross-validation iterations, producing an average per voxel, for that model. To validate this analysis procedure, we randomly selected one subject and performed the same regression and model evaluation procedure using 10 000 instances of scrambled labels on a random sample of 10 000 voxels. Across these iterations and voxels, the mean difference between correct (on-diagonal) and incorrect (off-diagonal) predictions when the regressions were performed with scrambled labels was 3.02×10^{-4} (median = 4.17×10^{-5}), close to the expected value of zero.

Finally, for conceptual clarity, we multiplied the average differences across iterations by −1 so that informative voxels are represented as greater than zero. A region whose learned encodings generalize to new sentences (have low prediction error) is thus presented as having a positive average difference between on-diagonal (matched) and off-diagonal (mismatched) prediction errors. For group-level analysis, these maps were then smoothed at 8 mm FHWM, warped to Talairach Space, and submitted to a two-tailed *t*-test against zero.

For all search analyses (both whole-brain and within Region-of-Interest (ROI)), we used clusterwise correction for multiple comparisons to control the familywise error (FWE) rate. To obtain these corrected *P* values, we used Monte Carlo simulations in AFNI (Cox 1996) (version 17.3.06). This simulation empirically estimates the probability of obtaining clusters of a certain voxelwise statistical magnitude and spatial extent, given that the data contain only noise. To estimate the smoothness of the noise, we randomly permuted the sentence labels for each subject and mimicked the individual and group procedures described above to obtain a group-level random statistical map, generated using the same procedures, but with noise-only data. We averaged 5 iterations of these noise-only group-level maps to obtain the spatial autocorrelation parameters for the Monte Carlo simulation. We used this procedure to correct across the whole-brain volume (216, 908 voxels), using a voxelwise threshold of $P < 0.005$, and a FWE of $P < 0.05$.

Whole-Brain Search

First, to identify regions potentially encoding complex, structure-dependent semantic representations throughout the entire brain, we evaluated voxels' ability to generalize to new sentences, using the full sentence model shown in Figure 2B. We call this the "full" model because it contains both broad and narrow predictor variables, as well as unstructured noun and verb variables. Thus, here, we seek to identify regions that carry any lexico-semantic information that generalizes across sentences and enables discrimination without specifying exactly what representations enable the prediction. This was used to localize an ROI, in which we pursue more targeted analyses below.

The full model included variables representing word identities (e.g., "hawk," "hog," "noticed"), recurring across sentences and semantic and syntactic roles (6 nouns + 8 verbs = 14 variables). The model also included variables encoding these nouns' interaction with other sentence components. These interaction terms allow the model to capture information that depends, not just on the stable semantic content of the words present, but also the way in which these words' meanings interact with others in the sentence, and with their assignment to particular structural positions. These included variables describing the nouns' interaction with particular verbs (e.g., "the hawk noticed") (6 nouns × 16 roles = 96 variables), particular grammatical positions (e.g., with hawk-as-subject) (6 nouns × 2 roles = 12 variables), particular classic semantic roles (e.g., hawk-as-experiencer) (6 nouns × 4 roles = 24 variables), and macro-roles (e.g., "hawk as the receiving entity in the event") (6 nouns × 2 roles = 12 variables). In total, this produces a model with 158 predictor variables and one intercept term. The predicted BOLD signal for a new sentence was modeled as a weighted, linear combination of these variables. However, for any voxel whose BOLD signal can be successfully predicted by the full model, we do not immediately know which parameters drive the success. We therefore did two follow-up analyses. First, we conducted a test of classification on mirror-order proposition pairs that contain the same words, but in which these words are combined to form different meanings ("the hawk approached the cow" vs. "the cow approached the hawk"). This enables us to focus on regions that encode relational information. Second, we partitioned the model into three separate, simpler models, enabling the identification of regions carrying information about specific nouns (bag-of-nouns), noun-verb combinations (narrow semantic roles), and abstract noun-role combinations (broad semantic roles), shared across verbs.

Mirror-Order Classification

We first asked whether any regions identified by the full sentence model could discriminate sentences that use the same words to create different meanings. We call these "mirror-order proposition pairs." This analysis was conducted as follows. We took, for example, the model's predictions for the propositions "the crow surprised the moose" and "the moose surprised the crow" (which could have been presented in either the active or passive voice). We generated model predictions for each of these two sentences, and asked whether the prediction for proposition one (e.g., "the crow surprised the moose") was more similar to the observed data than the prediction for proposition two ("the moose surprised the crow"), and vice versa (see Mitchell et al. 2008 for a similar evaluation procedure). We performed this classification for each of the 120 possible mirror-order pairs, averaging the results across pairs and across subjects. Finally, we conducted a two-tailed *t*-test against zero to assess whether the region identified by the whole-brain, full-model carried information discriminating mirror-order proposition pairs, as we would predict. For technical reasons, the amPFC ROI mask for one subject could not be properly inverted from Talairach space to his/her native space. The mirror-order classification analysis thus had 47 total subjects.

Analyses Using Component Models

For post hoc analyses (see Fig. 3b), we split the full model into three separate models targeting particular types of representations. The first model encoded only noun identity, invariant to the semantic role that noun occupied in a particular sentence. We call this model "bag-of-nouns". This model thus had only 6 parameters (one for each noun), plus a constant term. That is, this model only represents whether a particular noun was present, without respect to the role it played. The second model targeted only broad semantic role combinations (broad noun-role combinations), reused across verbs. Here, we employed 4 roles, for the agent/patient/stimulus/experiencer of the event. The particular roles for each verb were extracted from the role representations provided in the VerbNet database (Schuler (2005), http://verbs.colorado.edu/verb-index/index.php; see Methods section on Event and Sentence Structure for further discussion of psych verbs and these semantic roles.) This model treated each instance of a noun in these roles as identical across verbs. This broad model thus had 24 (6 nouns × 4 roles) parameters, and a constant term. The final model targeted verb-specific roles (narrow noun-role combinations). It thus had 6 nouns × 8 verbs × 2 = 96 distinct noun-role combinations. Although the narrow model has many more parameters than previous two models (96, 24, and 6 terms respectively), we note that the models are always evaluated on their ability to predict held-out data, testing their generalization ability using the same number of test trials.

For both the bag-of-nouns model and the broad role model, we modify the evaluation procedure slightly. Because these models, unlike the full model and narrow role model, ignore information about verbs, there exist cases in which the model will generate identical predictions to sentences with different meanings. For example, the broad-role model would generate the same prediction for "the moose approached the goose" and "the moose attacked the goose," as these contain the same nouns and the same broad noun-role combinations (moose-as-agent and goose-as-patient). Therefore, to ensure that we can directly compare the performance of these models to performance of the narrow model (which encounters no cases of identical prediction), we performed separate model evaluations for each trial while withholding these ambiguous cases for that iteration. Here, the model was still trained on 5 of 6 runs, as before. However, for each trial in the held-out run (40), any test trials that would generate identical predictions to that trial, using that model, were removed from model evaluation. Sentences that would generate similar, but not identical predictions were not withheld, as there is still broad semantic role information that the model could use to generate separate predictions. For example, if the test trial is "the moose approached the goose," the sentence "the moose approached the cow," would still be included, as the bag-of-nouns model and the broad role model generate different predictions based on the information present in the identity of patient. The prediction error *z*-scores were averaged across these 40 trial-iterations, for each run, and then submitted to the same group-level evaluation as the full model.

ROI Definitions

For the a priori lmSTC ROI, we used the union of the anterior agent and patient regions identified in Frankland and Greene (2015), down-sampled to 2 mm^3 voxels (see Fig. 3). This ROI is centered at (−53, −13, 0), (all anatomical coordinates are provided in Talairach space) and contains 275 2 mm^3 voxels. For the a priori hippocampal ROI, we used the "TT_desai_dd_mpm" anatomical atlas provided in AFNI. In this atlas, the subcortical masks, such as the hippocampal ROI employed here, are parcellated by FreeSurfer (see Fig. 3). This bilateral hippocampal ROI contained a total of 949 voxels (478 left hemisphere, 471 right hemisphere). The left hippocampal ROI was centered at (−22, −7, −11). The right hippocampal ROI was centered at (27, −18, −12). The amPFC ROI was identified by the whole-brain analysis described above, but here we defined the ROI shape using a more liberal voxelwise threshold ($P < 0.05$). It was centered at (−22, 54, 7), and contained 955 voxels.

Within-ROI Comparison of Encoding Models

What information is encoded in these three ROIs during sentence comprehension? To characterize their representational content, we performed the following two-step procedure. First, we searched these three small volumes (amPFC, lmSTC, hippocampus) using our three separate submodels (broad roles, narrow roles, and bag-of-nouns), assessing whether any clusters therein survived small-volume correction ($P < 0.01$ voxelwise, $P < 0.05$ clusterwise corrected). To determine whether these subregions differ significantly in their representational content, we then iteratively held each subject out of these searchlights, and selected clusters of voxels that exhibited $t > 2.4$, $P < 0.02$ with a volume of 40mm^3 in the remaining subjects. Within these independently localized regions for the held-out subject, we then averaged each model's performance. We repeated this cross-validation procedure for all 48 subjects, and performed a repeated measures ANOVA, testing for a "region × content" interaction. Such an interaction would demonstrate that voxel clusters within these ROIs, independently localized for each subject, differ significantly in their representational content. We then perform a series of planned pairwise *t*-tests evaluating the difference driving this interaction.

In addition to computing analytic *P* values based on the *F* and *t* distributions, we also perform permutation tests in which the region and model labels are randomly permuted 10 000 times and the statistics are again computed using these permuted labels. We use this as the null distribution to compute the permutation *P* values, reflecting the probability of obtaining the true statistic under random permutation. Although these yield similar statistics, we report both throughout the paper.

Similarity and Structure: ROI analyses

The analyses noted thus far treat each noun and verb as fully distinct variables, ignoring similarity structure based on shared features. That is, the nouns are six discrete variables ignoring phonological and semantic similarity relations that may exist between them. Here, we exploit two key design features of our stimuli in follow-up analyses. First, the sentences contain a class of "psych" verbs enabling the dissociation of the semantic structure of the event from the syntactic structure of the sentences (see Fig. 4a). Second, the nouns cluster into dissociable semantic and phonological similarity spaces (see Fig. 2d). Both of these properties enable us to better study the representational content of our regions of interest.

Semantic and Syntactic Structure of the Verbs

Stimuli

First, we describe the analyses concerning psych verbs. We treat the 8 verbs we employed as falling into one of two broad sets. Set 1 contains the verbs 'chased', 'approached', 'passed', and 'attacked'. These verbs are characterized by different manners of motion/intention of the agent with respect to the patient. Set 2 consists of 4 verbs that additionally conveyed some aspect of a mental state, referred to as "psych verbs." The four psych verbs could then be further subdivided into two groups of two, the members of which were internally similar both semantically and syntactically: (Set 2a: 'surprised', 'frightened'), and (Set 2b: 'noticed', 'detected'). In the first subclass (Set 2a), the participant undergoing the experience ("the experiencer") is the object of the active-voice sentence, while the participant causing or serving as the content of the experience is the subject ("the stimulus"). We refer to Set 2a as "experiencer-object" sentences. By contrast, for Set 2b, the event participant undergoing the experience is the subject ("experiencer-subject") of the active-voice sentence, while the stimulus is the object. Thus, these

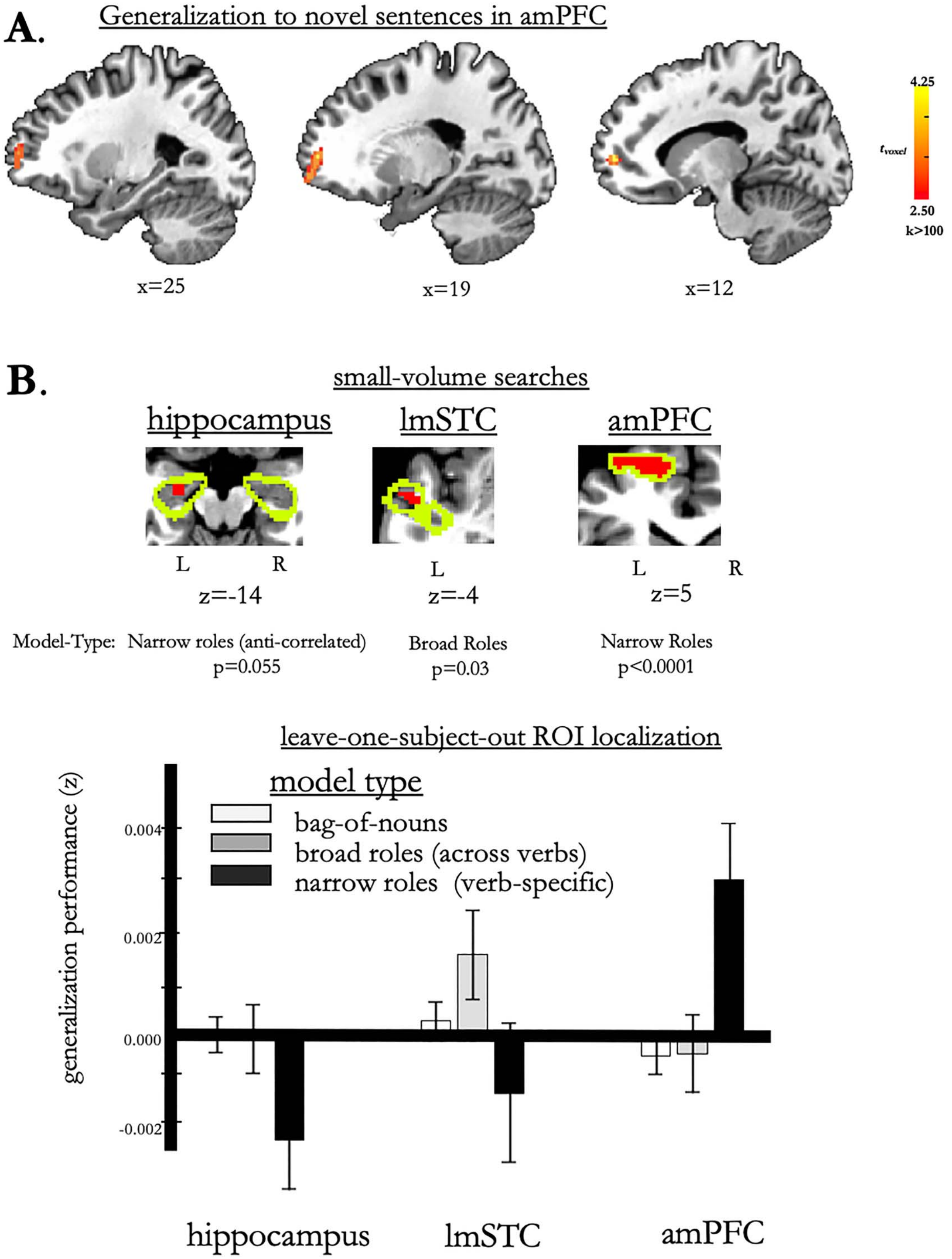

Figure 3. Encoding models reveal that different brain regions use distinct strategies for representing who did what to whom (*A*) Our full encoding model identifies a significant cluster in anterior-medial prefrontal cortex (BA10, (peak, −13, 53, 6)), ($P < 0.005$ voxelwise, $k = 203$, $P = 0.0001$, whole-brain corrected) in which learned model parameters predict significant variation in BOLD signal on held-out, novel sentences. (*B*) We split the full encoding model into three submodels reflecting different representational strategies. Across three ROIs, we compare these sub-models' ability to predict BOLD signal to novel sentences. One model uses terms indicating the presence of specific nouns, independent of their semantic roles and the present verb (bag-of-nouns—e.g., "cow" appears in the sentence). A second model uses terms for nouns bound to abstract event-roles, which also generalize across verbs (broad roles—e.g., "cow" is the agent in the sentence). A third model uses terms for nouns in combinations with specific verbs (narrow roles—e.g., "cow" is the entity that "chases"). These three encoding models show different patterns of performance across these three regions (green outline), identifying significant subregions (red) that represent information about who did what to whom in distinct and complementary ways. Bars in the plot represent average model performance in the red-regions, defined for each subject using independent data from the other subjects. There is a significant encoding model × region interaction ($F(4188) = 7.84$, $P = 7.18 \times 10^{-6}$). Error bars reflect standard error of the mean.

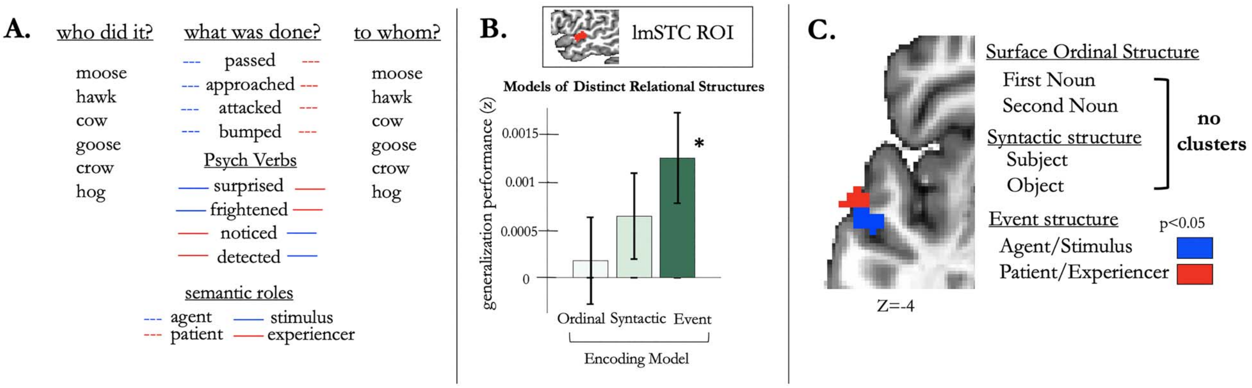

Figure 4. Representation of event structure in lmSTC. Panel (A) shows the mapping from active-voice sentence structure to semantic roles (solid and dashed colored lines). Critically, in addition to standard agent/patient verbs, four of the 8 verbs referred to events conveying a change in the participant's psychological state. This class of verbs, known as "psych verbs," is unique in that it allows for dissociation of the sentence syntax (subject/object), from the semantic role in the event. By "event structure," here, we refer to the causal/temporal structure of the event: the entity that causes the psychological event (the "stimulus") is grouped with the agent of other verbs (e.g., the attacker), and the entity undergoing a change of psychological state (the "experiencer") is grouped with the patient (e.g., the attackee). (B) Models based on this grouping by event structure explain significant variance in lmSTC, but models based on subject/object groupings alone and ordinal structure ("surface syntax") does not. ∗Statistically significant ($P < 0.05$) generalization performance (C).Within the lmSTC ROI, we also find significant clusters for the agent/stimulus ($k = 29$, $P = 0.015$ clusterwise) and patient/experiencer groupings ($k = 18$, $P = 0.042$ clusterwise), but none for individual roles based on syntactic or ordinal structure.

sentences involving psych verbs dissociate deep subject/object (syntax) relations from semantically coarse stimulus/experiencer relations (Pesetsky 1987; Belletti and Rizzi 1988; Dowty 1991; Levin and Hovav 2005). In our previous work (Frankland and Greene 2015), we searched for neural representations that generalized across active and passive forms (surface syntax). However, there, we could not differentiate deeper syntax from event structure, given their close relationship. Psych verbs offer a unique opportunity to not only differentiate surface from deep syntactic structure ("the hawk frightened the moose" from "the moose was frightened by the hawk"), but also deep syntactic structure from event structure. Compare the sentences "the hawk frightened the moose" and "the hawk noticed the moose." In the first, the verb "frightened" refers to an experiential change in the moose (which, in this sentence, is 1) the surface object, 2) also the "deep" object, as the sentence is in the active voice, and 3) the undergoer of the experience in the event described. By contrast, in the sentence "the hawk noticed the moose" the experiential change is undergone by the hawk (here, 1) the surface subject, 2) the deep subject, but still the 3) undergoer of the experience. We can therefore ask whether a particular brain region groups sentences by syntactic or semantic invariances (causal event structure), providing key information as to its representational function.

Analyses

We evaluated the predictive performance of models organized around three groupings of the nouns (see Fig. 4). The first grouping is by the ordinal structure of the noun presentation ("surface syntax"). This model would, for example, treat the sentences "the hawk approached the cow" and "the hawk was surprised by the cow" as identical, as, reading from left to right, "hawk" is presented first, and "cow" second, despite the fact that they have different roles in the underlying syntax and semantics. The second organizes nouns by the deeper subject/object relations ("underlying syntax"). This model would, for example, group "the hawk approached the cow" with "the cow was surprised by the hawk," as the "hawk" is the underlying subject of the active construction, and "cow" the object, despite the difference in order. The third model groups stimuli by the underlying event structure. This is closely related to subject/object grouping, but differs critically for the psych verbs. Here, we group the stimuli based on the causal-temporal structure of the event described. Grouping by this aspect of event structure clusters the agent and stimulus variables together, and the patient and experiencer verbs together. As an illustration, consider the events described by the verbs "bumped," "surprised" (experiencer-object), and "noticed" (experiencer-subject). The entity that "does the bumping," as well as the entity that "does the surprising" initiate the event, and would thus be considered causally responsible. Likewise, their action is temporally prior to the change of state encoded by the verb. These are canonically mapped to subject position. However, for experiencer-subject verbs, such as "noticed," the entity that notices (the subject) observes some state that existed prior to the noticing event. Thus, here, the subject, is affected, like the object of "surprised" and the object of "bumped." Thus, we group agent with stimulus, and patient with experiencer based on the causal structure of the event. This creates semantic roles existing at a broader level of description than classic categories like agent and patient (see Dowty 1991; Van Valin Jr and Van Valin 2005; Levin and Hovav 2005). Although there may good reason to additionally expect narrower semantic roles that further differentiate the experiencer of experiencer-object and experiencer-subject sentences (and likewise for the stimulus) (Belletti and Rizzi 1988; Hartshorne et al. 2016; Ziegler and Snedeker 2018), here, we focus on the abstract semantic features that are shared. For example, this model would group "the hawk approached the cow" with "the cow noticed the hawk," as the hawk is considered the agent/stimulus of the event, and the cow the patient/experiencer.

Similarity Structure of the Nouns

Stimuli

The particular nouns were chosen because their phonological similarity relationships are distinct from their semantic similarity relationships, enabling us to study the encoding scheme of identified regions. Semantically, all nouns refer to animal categories, and naturally divide into a group of "mammals" ('moose', ‘cow’, ‘hog’), and a group of “birds” (“crow,” “goose,” ‘hawk’). Phonologically, each noun in one semantic class (e.g., ‘hog’ in the mammal class) has a strong phonological associate in the other semantic class (e.g., ‘hawk’ in the bird class). This distinction is particularly relevant to understanding the lmSTC ROI, as nearby regions of lmSTC have been reported to respond to both phonological (Vigneau et al. 2006; Poeppel et al. 2004; Belin et al. 2002) and semantic (Vigneau et al. 2006; Rodd et al. 2015; Price et al. 1997) manipulations in functional neuroimaging. In one meta-analysis (Vigneau et al. 2006), lmSTC was the only region in which phonological, semantic, and syntactic contrasts overlapped significantly in the imaging studies reviewed. It is thus possible that the regions we study here could represent words, and do so using a phonological code that treated similar sounding words as similar. Or, they could represent noun-meanings, using a semantic code that treated similar objects as similar. Failure to support either of these models leaves open an intriguing third possibility: that these regions do not use similarity-based coding schemes, but instead use an effectively arbitrary coding scheme.

In quantifying phonological similarity, we were guided by Mueller et al.’s (2003) PSIMETRICA model of phonological similarity. The phonetic constitution of a syllable can consist of three sequential parts: an “onset” (0–3 phonemes), a “nucleus” (1–2 vowel phonemes) and a “coda” (0–5 consonant phonemes). The nucleus and coda jointly constitute the “rhyme.” Our stimuli consist of three pairs of phonologically similar nouns. Each phonologically similar pair is similar along a subset of Mueller et al.’s, dimensions: similar onset and nucleus {‘hog’,‘hawk’}, similar onset {‘crow’,‘cow’}, or similar nucleus and coda {‘goose’,‘moose’}.

Analyses

To explore the similarity structure of these regions, we re-ran models that organized the nouns according to their semantic and phonological structure. The semantic model simply grouped nouns by into the category of “mammal” or “bird.” The phonological category group similar pairs together (e.g., “moose/goose”) (see Fig. 2D for similarity spaces).

amPFC-Hippocampal Beta-Series Connectivity

Given previous work on the functional dependence between hippocampus and medial prefrontal regions in spatial and episodic memory tasks (e.g., Zeithamova et al. 2012), we asked whether generalization ability in amPFC predicts generalization ability in the hippocampus. To perform this analysis, we mapped the a priori ROIs (defined above) from Talairach space to each subject’s native space. We then trained the full models within these ROIs, as described above, and then tested for a correlation between the trial-by-trial prediction errors. To assess the statistical significance of trial-by-trial performance, we conducted a second-level t-test assessing whether the correlation coefficients between regional prediction errors were significantly nonzero across subjects. As with the mirror-order classification analysis, one of the 48 subjects was excluded from this analysis because his/her ROIs could not be reliably mapped from Talaraich space to native space, leaving 47 subjects for this particular analysis.

Conceptually, this time series analysis is closely related to beta-series connectivity analyses (Rissman et al. 2004), as we are comparing the model-fits across regions over time. Here, we focus on such time series correlations using the model fits of rich encoding models. Our analysis thus also bears a conceptual relationship to the class of “informational connectivity” analyses (cf. Coutanche and Thompson-Schill 2013; Frankland and Greene 2015; Anzellotti and Coutanche 2018), in that we are interested in the stimulus-dependent synchrony between two information-bearing states over time.

Results

Whole-Brain Search Results

Across subjects, this analysis using the full model (Fig. 2b) revealed a significant cluster of voxels ($P < 0.005$ voxelwise, $k = 203$, $P = 0.0001$, whole-brain corrected) in anterior-medial prefrontal cortex (amPFC) (medial frontal gyrus, BA10) in which learned model parameters predict significant variation in BOLD signal across novel sentences (see Fig. 3a). The region is left-lateralized, adjacent to the midline, and centered at (−22, 54, 7, Talairach space, peak: −13, 53, 6). This is the only cluster that survived whole-brain correction.

amPFC Mirror-Order Classification

Our primary goal is to understand the brain’s strategies for dynamically encoding the structured relations in an event. However, given that the full model contains variables for “unstructured” nouns, the whole-brain search result could be driven by the mere presence of a noun, as would be predicted by a bag-of-words model, commonly used as baseline models in computational linguistics. Related unstructured models have been used to predict neural activity in other brain regions (Anderson et al. 2016). Given that our primary interest is in structured semantic composition (who did what to whom), we sought to determine whether amPFC’s generalization to new sentences owes to structure-dependent or structure-independent representations. To do so, we first asked whether amPFC patterns can discriminate sentences that contain the same words, but express different relations between the event participants using mirror-order proposition pairs (e.g., the crow surprised the moose vs. the moose surprised the crow). Indeed, across subjects, the full amPFC model reliably discriminated mirror-order proposition pairs ($t(47) = 2.6$, $P = 0.012$), providing evidence that it carries structured (i.e., relational) information, sensitive to the roles played by the event participants. We next sought to determine the level of abstraction (broad vs. narrow roles) and also compared the representational profile of the amPFC ROI to two a priori ROIs.

Representational Profiles of amPFC, lmSTC, and Hippocampus

Within ROI Search

All within-ROI search results were corrected for multiple comparisons, using clusterwise correction, as in the whole-brain search, but within a small-volume. Within the amPFC ROI, we find a cluster of voxels whose activity is predicted by the narrow role model. This cluster constitutes the entire ROI localized using the whole-brain search ($P < 0.00001$, $k = 203$ of 203 in ROI); however, no such clusters were found in amPFC for the broad-role model or the bag-of-nouns model. By contrast, within lmSTC, only the broad role model yields a significant cluster ($P = 0.03$, $k = 13$). lmSTC contains no significant clusters for the narrow model or bag-of-nouns model. Moreover, the hippocampus shows a different pattern than either amPFC or lmSTC. Here, we see a marginally significant negative effect for the narrow model ($P = 0.055$ small volume corrected, $k = 18$) within a left-anterior portion of the hippocampus, small-volume corrected within the anatomically defined bilateral hippocampal ROI. Within this hippocampal cluster, sentences that share narrow noun-role representations but that otherwise differ (e.g., "the moose surprised the hawk" vs. "the moose surprised the cow") are more dissimilar to one another than those that do not share representations. We find no significant clusters for broad or bag-of-nouns models in hippocampus.

Post Hoc Analysis of Representational Content across ROIs

The differential performance of distinct encoding models across our three main ROIs suggest that amPFC, lmSTC, and the hippocampus make different contributions to the compositional representation of complex events. Here, we evaluate these differences more directly, testing for statistical interactions between region and the performance of distinct encoding models. We separately localized the above three regions using N-1 subjects, and averaged the performance of each model for the held-out subject within each ROI. This cross-validation analysis reveals a statistically significant interaction ($F(4188) = 7.84$, $P = 7.18 \times 10^{-6}$, $P_{\text{perm}} = 7.2 \times 10^{-5}$), confirming that these subregions differ significantly in their representational content (see Fig. 3b.)

Consistent with the results of our search within the amPFC ROI, we find that only the narrow role model predicts activity in response to novel sentences in amPFC ($t(47) = 2.76$, $P = 0.008$, $P_{\text{perm}} = 0.0078$). The broad-role ($t(47) = -0.46$, $P = 0.64$, $P_{\text{perm}} = 0.65$) and bag-of-nouns ($t(47) = -1.08$, $P = 0.28$, $P_{\text{perm}} = 0.28$) do not predict responses to held-out sentences, and indeed are significantly worse than the narrow role model (narrow > bag-of-nouns: $\mathrm{t}(47) = 3.29$, $P = 0.002$, $P_{\text{perm}} = 0.0017$. narrow > broad: $t(47) = 2.75$, $P = 0.0085$, $P_{\text{perm}} = 0.0087$). This narrow-role model's performance in amPFC is significantly greater than its performance in both the identified lmSTC ($t(47) = 2.86$, $P = 0.006$, $P_{\text{perm}} = 0.0061$) and hippocampal subregions ($t(47) = 3.87$, $P = 3.36 \times 10^{-4}$, $P_{\text{perm}} = 4 \times 10^{-4,}$). By contrast, this lmSTC subregion carries no information about narrow noun-role combinations ($\mathrm{t}(47) = -0.82$, $P = 0.41$, $P_{\text{perm}} = 0.41$). Instead, we see a trend toward significant broad-role generalization across subjects ($\mathrm{t}(47) = 1.80$, $P = 0.078$, $P_{\text{perm}} = 0.077$), a nonsignificant trend toward greater performance on the broad model in lmSTC than the broad role model in amPFC ($t = 1.68$, $P = 0.098$, $P_{\text{perm}} = 0.097$), and significantly greater performance than the broad role model in the hippocampal subregion ($t(47) = 2.16$, $P = 0.035$, $P_{\text{perm}} = 0.034$). Within our lmSTC region, there is a marginal effect of better performance for the broad role than narrow role model ($t(47) = 1.90$, $P = 0.063$, $P_{\text{perm}} = 0.062$), showing the opposite effect as amPFC, and no significant effect of bag-of-nouns ($t(47) = 0.73$, $P = 0.47$, $P_{\text{perm}} = 0.47$). (We further evaluate the particular representational content of lmSTC in the section below titled "Event Structure, Syntactic Structure, and Ordinal Structure within the ROIs.") Finally, the anterior hippocampal ROI is significantly "below chance" at predicting narrow role combinations ($t(47) = 2.10$, $P = 0.04$, $P_{\text{perm}} = 0.039$), but not significantly different from zero using either broad roles ($t(47) = -0.09$, $P = 0.92$, $P_{\text{perm}} = 0.93$) or bag-of-nouns models ($t(47) = 0.03$, $P = 0.97$, $P_{\text{perm}} = 0.97$). Direct comparisons reveal that the narrow role model in this hippocampal ROI is significantly worse than bag-of-nouns models ($t(48) = -2.26$, $P = 0.028$, $P_{\text{perm}} = 0.027$), and is marginally significantly worse with respect to broad roles ($\mathrm{t}(47) = -1.96$, $P = 0.056$, $P_{\text{perm}} = 0.052$). We note that, unlike our searchlights, these post hoc t-tests are reported uncorrected for multiple comparisons across tests. However, taken in conjunction with our searchlight results, the pattern of results strongly suggests 1) that the identified region of BA10 (amPFC) encodes narrow noun-role conjunctions 2) an anterior portion of the left hippocampus shows the opposite effect, exhibiting below-chance generalization performance, and 3) lmSTC represents more abstract roles than amPFC and the hippocampus, ignoring verb-specific information in favor of broader role representations.

Event Structure, Syntactic Structure, and Ordinal Structure within the ROIs

The foregoing models targeted the representation of semantic relations by treating active and passive constructions involving the same semantic structures as equivalent. Each proposition was randomly presented in the either active or passive voice, with different randomizations across subjects. Here, we focus on the differences between related types of structure, targeting differences in event structure, syntactic structure, and ordinal structure more directly (note, here, we use the family of terms surrounding "semantic representations" and "conceptual representations of event structure" interchangeably, as is standard in psychology, but not linguistics. We acknowledge that "semantic" may ultimately deserve a narrower construal tied to lexical meaning, but, here, keep with standard practice in our field.)

Event Structure in lmSTC

We can begin to tease apart event representation and syntactic representation by exploiting the inclusion of psych verbs in the stimulus set, in which the mapping between semantic roles (event structure) and syntactic roles varies between experiencer-subject (e.g., "noticed") and experiencer-object (e.g., "surprised") verbs. To evaluate event vs. syntactic structure, we carve our broad role model into two lower-dimensional representations. One captures the underlying syntactic structure of each sentence, grouping together the first noun of the active-voice construction ("the moose [did something to something]") and the second noun of the passive construction ("[something had something done to it] by the moose"), and likewise grouping the active-voice second noun with the passive voice first noun. The other captures the semantic structure of the event (e.g., grouping the subject of the active-voice construction of "noticed" with the object of the active-voice construction of "surprised," as both reflect the "experiencer" role). We group the agent and stimulus roles together and patient and experiencer roles together to reflect the causal-temporal structure of the event, thus subsuming classic thematic roles (see Dowty 1991; Van Valin Jr and Van Valin (2005) for related abstract "macro-role" models in linguistics). Both predictive models here, syntactic and semantic, thus had 2 roles × 6 nouns to generate 12 parameters, plus a constant term.

We find that the low-dimensional semantic role model (agent or stimulus/patient or experiencer) predicts significant variation within lmSTC ($\mathrm{t}(47) = 2.64$, $P = 0.01$, $P_{\text{perm}} = 0.016$), while the syntactic role model (deep subject/object grouping) does not significantly predict neural activity therein ($t(47) = 1.42$, $P = 0.16$, $P_{\text{perm}} = 0.16$). Although the direct comparison of these two models is not statistically significant ($t(47) = 1.51$, $P = 0.14$, $P_{\text{perm}} = 0.14$), this difference is notable, given that these models vary only in their encoding of experiencer-subject psych verbs: for example, for the event structure model, the "noticer" (subject) is grouped with the grammatical object of the verb “surprised,” while the “noticed” (object) is grouped with the grammatical subject of “surprised.”

This causal-temporal mapping is not the only partition of the 4 roles into more abstract groups, however. For example, an alternative mapping might sort roles by “mental state” information, grouping the agent and experiencer (and patient with stimulus) under the assumption that mental state information is more important for the agent and experiencer than the stimulus and patient. However, this model is also nonsignificant in lmSTC ($t(47) = 1.61$, $P = 0.11$, $P_{\text{perm}} = 0.112$). Finally, as expected, a model that only encodes the surface ordinal structure of the nouns (rather than deep syntactic structure) does not predict lmSTC activity ($t(47) = 0.38$, $P = 0.71$, $P_{\text{perm}} = 0.70$). Thus, we find that a model that encodes semantic rather than syntactic or ordinal structure best predicts BOLD signal to novel sentences in lmSTC. Moreover, an event structure model that respects that causal-temporal structure of the event, grouping agent(stimulus) and patient(experiencer) appears most promising.

Agent/Patient/Event Organization

For the models discussed so far, a voxel’s activity is predicted as a function of the learned weights for multiple roles: for example, exploiting both the agent and patient variables to model activity in a single voxel. However, in previous work (Frankland and Greene 2015), we found dissociable subregions of lmSTC that differentially contribute to the representation of the agent (medial) and patient (lateral) of the event. We therefore broke the semantic and syntactic models into smaller role-specific models (e.g., agent, patient separately), and asked whether any clusters within this lmSTC region could predict activity to held-out sentences using particular roles alone. Within lmSTC, we find a sub-region in which the identity of the causal entity (agent/stimulus) is encoded ($P < 0.05$ voxelwise, $k = 29$, $P = 0.015$ clusterwise), as well as a marginally significant adjacent sub-region in which the identity of the causally affected entity (patient/experiencer) is encoded ($P < 0.05$ voxelwise, $k = 18$, $P = 0.042$, clusterwise; see Fig. 4). By contrast, we find no clusters in lmSTC tracking underlying syntactic structure (subject/object) or surface structure (first/second). This result provides further evidence that lmSTC is representing abstract event structure rather than syntactic roles.

Notably, these clusters share the same topographic organization observed in Frankland and Greene (2015) in which the medial region carries information about agent (here, agent and stimulus) while the lateral carries information about the patient (here, patient and experiencer; see Fig. 4c). However, unlike previous work, we do not find evidence here for a region-X-content interaction between the immediately adjacent medial/lateral portions of STG ($F(4,88) = 0.14$, $P = 0.71$) and thus the present results do not provide evidence for the stronger claim that these are role-selective regions. However, taken collectively, these analyses provide evidence that an anterior portion of lmSTC supports reusable semantic representations that combine to encode aspects of event structure, rather than surface or deep syntactic features.

The analyses presented thus far provide evidence that lmSTC encodes relational combinations at a greater level of abstraction than amPFC. However, in order to encode the entire proposition (to have the full set of materials necessary to “build a thought”) the identity of the event-type must also be preserved in some form. That is, one must know not just “who did it?” and “to whom was it done?”, but also “what was done?”. We thus also evaluated whether any regions of the left temporal lobe could be predicted by a verb-identity model that captures the trial-by-trial identity of the verb, invariant to the nouns with which it co-occurs. This verb-identity model thus had 8 parameters (one for each verb) and a constant term. As we would not expect such a region to anatomically coincide with the verb-invariant agent/patient regions, we searched a large left temporal ROI (11 115 voxels) for regions carrying information about verb identity. This analysis revealed a significant verb-identity cluster in left STG that generalized across sentence contexts ($k = 83$, $P = 0.015$ clusterwise corrected, center = −49, 10, −5) to predict the trial-by-trial identity of the verb (that is, the event-type). This cluster is near, but anterior to, the agent/patient regions (see Supplementary Fig. 2). By contrast, neither the amPFC ROI, nor the hippocampal ROI contain any significant verb-identity clusters. Critically, this is not to say aMPFC and hippocampus are insensitive to the trial-by-trial identity of the verb, but only that the form of the representation differs between these regions and lmSTC. Recall, we find that amPFC and the hippocampal ROI are predicted by the narrow role model, which integrates information about the verb and noun occupying the role into one conjunctive representation (moose-as-approacher), rather than keeping these representational components separate. Moreover, we do not suggest that STG is the only brain region capable of identifying trial-by-trial abstract verb identity. However, the fact that STG does contain such a representation is consistent with a highly general account of its potential representational contribution: specifically, that it favors low-dimensional representations of recurring aspects of event structure, separately encoding abstract factors such as “who did it,” “to whom was it done” and “what was done.”

Semantic vs. Ordinal Structure in amPFC and hippocampus

amPFC and the left-anterior hippocampal ROI exhibit significant effects for narrow-role models, rather than broad-role models (as seen in lmSTC). However, we note that a region could also potentially represent ordinal relations among words narrowly or broadly, just as one can represent semantic relations narrowly or broadly (as presented in Fig. 1). For example, compare ordinal relations bound to a particular verb, such as “cow-before-approached” to ordinal relations that are invariant across particular verbs “cow-as-first-noun.” Here, we compare performance of narrow semantic roles to narrow-ordinal roles in amPFC and hippocampus. Using the leave-one-subject-out procedure described above, we find that models encoding narrow-ordinal structure in the sentence do not predict BOLD signal to novel sentences within amPFC ($t(47) = 0.94$, $P = 0.35$, $P_{\text{perm}} = 0.35$) or the left-anterior hippocampus ($t(47) = 0.13$, $P = 0.90$, $P_{\text{perm}} = 0.90$). In direct comparison, the amPFC ROI trends toward better predictive performance when using narrow semantic rather than narrow-ordinal relations ($t(47) = 1.89$, $P = 0.065$, $P_{\text{perm}} = 0.063$). By contrast, the left-anterior hippocampus trends in the opposite direction ($t(47) = -1.70$, $P = 0.095$, $P_{\text{perm}} = 0.096$). However, this negative effect is due to the reliably below-chance performance in this hippocampus ROI for semantic relations, as the ordinal model shows no trends in either direction ($P = 0.90$), in contrast to the narrow semantic relation model ($P = 0.04$ for a negative effect). Though we ensure statistical independence by localizing the ROI for each subject separately, it is still perhaps unsurprising that these particular regions are sensitive to the reuse of semantic relations. Thus, for completeness, we further searched the broader amPFC and bilateral hippocampal ROIs to ask whether other clusters might encode ordinal rather than semantic relations. Here, we found no significant clusters for the narrow-ordinal models in either amPFC or the bilateral hippocampal ROI. Note, we do not mean to suggest that such representations do not exist elsewhere in the brain (see Dehaene et al. 2015), but only that, in the current context, the particular regions of focus are sensitive to relational aspects of semantic variation, rather than superficial ordinal variation.

Similarity Structure of Nouns in These ROIs

Because the nouns in our stimulus set vary systematically in their semantic and phonological properties (see Fig. 2d), we ask whether the representations identified in the amPFC and elsewhere reflect semantic similarity (e.g. "hog" similar to "cow"), phonological similarity (e.g. "hog" similar to "hawk"), both, or neither. Within the entire amPFC, lmSTC, and hippocampal regions identified earlier, we find no evidence for representation of semantic similarity or phonological similarity (lmSTC: semantic, $t(47) = 1.5341$, $P = 0.13$, phonological, $t(47) = -0.6331$, $P = 0.53$. amPFC: semantic, $t(47) = 1.2072$, $P = 0.23$ phonological $t(47) = 0.0962$, $P = 0.93$, hippocampus: semantic, $t(47) = 1.39$, $P = 0.17$, phonological, $t(47) = 0.4861$, $P = 0.63$). However, all three regions trend toward effects of semantic similarity. For further exploration of regions outside these ROIs that track semantic and phonological similarity among these nouns, see Supplementary Materials.

amPFC-Hippocampal Beta-Series Connectivity

We find a significant negative relationship between amPFC and hippocampus model performance ($t(46) = -6.05$, $P = 2.37 \times 10^{-7}$). That is, the better the model predicts activity on a particular held out sentence in amPFC (lower prediction error), the worse it predicts hippocampal activity on that sentence (greater prediction error), and vice versa. This anticorrelation between amPFC and hippocampus is consistent with the representational profile identified above, wherein amPFC reuses narrow role representations across sentences, and an anterior sub-region of left hippocampus separates sentences sharing narrow conjunctions.

Discussion

Understanding how the human brain builds complex meanings out of simpler ones is a central problem for cognitive neuroscience (see Pylkkänen 2019; Frankland and Greene 2019 for recent reviews). Here, we have focused on a particular type of complex meaning (propositions involving an agent, patient, and event-type) that requires flexibly and dynamically encoding the relations among the event's participants. Using competing encoding models, we show that regions within the frontal and temporal lobes both carry information about who did what to whom in the event, but vary in the level of abstraction in the relations they encode.

amPFC and the Reuse of Conceptual Conjunctions

Using a whole-brain search, we identified a region of anterior-medial prefrontal cortex (amPFC, BA10) whose learned representations generalize to new sentences (Fig. 3a). This region distinguishes between members of sentence pairs that contain the same elements in different relational configurations (e.g. "the cow approached the goose" vs. "the goose approached the cow," across active and passive voice). Comparing encoding models, we found no evidence that amPFC's generalization owes to simply registering the presence of particular nouns (bag-of-nouns model), irrespective of their role. Nor did we find evidence that the amPFC makes use of broad roles shared across verbs. Instead, we found evidence that the amPFC uses narrow roles, such that representations of specific noun-verb conjunctions are re-used across sentences. For example, a representation of cow-as-approacher (encoded differently from cow-as-approachee) is reused in both "the cow approached the goose" and "the cow approached the hawk." That is, within amPFC, an encoding model based on specific, structured noun-verb conjunctions successfully predicted activity associated with new sentences. (Fig. 3b). This could not be explained simply by the ordinal structure of the words in the sentence (i.e., surface syntax), as, for example, "the cow approached" and "was approached by the cow" were treated equivalently in the narrow-role model. Moreover, alternative models that instead encode such superficial ordinal structure (e.g., treating "approached the cow" and "approached by the cow" equivalently) do not significantly predict amPFC activity. In sum, the amPFC is not encoding specific strings of words, but is instead encoding the underlying semantic relations, expressed by noun-verb combinations.

Although the mPFC is widely implicated in valuation and decision-making, the present findings are consistent with its representational role in memory-dependent tasks (Binder et al. 2009; Kumaran et al. 2009; Zeithamova et al. 2012; Preston and Eichenbaum 2013), including semantic inference (Pylkkänen and McElree 2007; Pylkkänen 2008), and the composition of complex concepts (Graves et al. 2010; Bemis and Pylkkänen 2011; Barron et al. 2013). For example, Bemis & Pylkkänen (2011) report MEG evidence that combining words such as "red" and "boat" to form "red boat" produces greater mPFC activity than noncompositional word pairs ("cup," and "boat"). Here, we use subject-verb-object combinations rather than adjective-noun combinations. However, both cases require computing the interaction of the constituent concepts. How do you modify a boat so as to make it red? How do you modify a chasing event so as to make a hawk the thing doing the chasing? Critically, we find that these complex representations are themselves reused across semantic contexts, supporting generalization to unfamiliar sentences.

We note that these findings dovetail with work on spatial and episodic memory that suggests that mPFC reuses knowledge structures to represent novel combinations of familiar components (Tse et al. 2007; Tse et al. 2011; Zeithamova et al. 2012; Preston and Eichenbaum 2013). For example, in rodents, mPFC promotes rapid (one-shot) learning of new food-location combinations, when the rodent has an intact mPFC and a pre-existing representation of the space. In humans, mPFC activity during associative learning of repeated object-image pairs (AB, BC) predicts inference of novel associations at a later time (grouping A and C) (Zeithamova et al. 2012). Notably, the abstract computational demands imposed by these memory paradigms share features with the composition of novel sentence meanings, as both involve reusing representations in particular relational configurations to interpret a novel combination.

In situating our results within these literatures, it is, however, important to note mPFC's anatomical heterogeneity. The particular portion that we identify is largely left-lateralized, and lies along the medial frontal gyrus (BA 10) at the frontal pole. The specific location of this cluster likely differs slightly from the precise location of previous effects that have implicated mPFC in conceptual knowledge acquisition (Kumaran et al. 2009), conceptual combination (Bemis and Pylkkänen 2011; Barron et al. 2013), associative inference (Zeithamova et al. 2012), and semantic coercion (Pylkkänen and McElree 2007), which tend to be adjacent to BA10, but more medial than the present region and more posterior. Ramnani & Owen (2004) suggest that BA10's common function across tasks is the integration of separate representations in pursuit of a goal. In the present paradigm, this integration may operate over event-types ("chasing") and the nouns playing particular roles in the event ("hawk"). Here, we provide evidence that amPFC's representation is not just limited to associations, but is also sensitive to the role the entity plays in the event (e.g, differentiating "the woman chased" from "chased the woman"). This relational representation is consistent with previous work identifying a role for polar regions of PFC in relational reasoning (Bunge et al. 2009; Knowlton et al. 2012) including analogy (Volle et al. 2010; Green et al. 2009; Urbanski et al. 2016). Moreover, we find that amPFC reuses these representational components across sentences that share the conjunction. Taken collectively, this work suggests that mPFC plays a critical role in reusing extant structured knowledge to encode unfamiliar, relational combinations.

lmSTC and the Representation of Abstract Event Structure

Using narrow roles is not the only way to represent relational event structure. Building on prior work (Frankland and Greene 2015), we also examined the performance of the same three encoding models in an a priori region of lmSTC, previously found to carry information about who did what to whom in an event. Here, too, we found no support for a bag-of-nouns encoding model, whereby patterns of activity reflect the presence of specific nouns, independent of their relations to other semantic elements. Critically, and in contrast to findings for amPFC, we also found no evidence for narrow noun-verb combinations. Instead, we find evidence for the representation of broad noun-role combinations that generalize across particular verbs (Fig. 3b, center). The statistical interaction between encoding models and regions (see Fig. 3) demonstrates that the distinctive success of the narrow role in amPFC (narrow > broad) cannot be explained by generic differences between these two models. This double dissociation between successful model type and region indicates that these regions represent semantic relations at different levels of abstraction. Moreover, we find that a nearby region of left STG (but not amPFC) carries information about the trial-by-trial identity of the verb, generalizing across sentence contexts. lmSTC thus appears to carry all the basic structural pieces necessary to begin to reason about a sentence's unique meaning—at least for simple sentences of the kind examined here.

This lmSTC effect is broadly consistent with the results of Frankland and Greene (2015), which provided evidence that distinct subregions of the anterior portion of lmSTC differentially encode the identity of the agent and patient. However, from previous work, it was unclear whether these broad roles are semantic representations of event structure, or syntactic representations of the underlying sentence structure, as these are tightly coupled. Here, we used psych verbs to begin to de-confound these variables. We find that lmSTC is better explained by the event structure, grouping the experiencer with the patient of other sentences, and the stimulus with the agent, regardless of underlying syntax, yielding significant effects in a more medial region (decoding the identities of agent/stimulus) and in a more lateral region (decoding the identities of the patient/experiencer) in a more lateral region. These medial and lateral regions correspond to the agent and patient regions identified previously (Frankland and Greene 2015). We find no such results in lmSTC when the sentences are grouped by underlying syntax (deep subject/deep object) or superficial ordinal structure of the nouns (surface subject/surface object). We note, however, that, here we do not find evidence that these medial and lateral portions of lmSTC are significantly different from one another, as we did in our original study (Frankland and Greene 2015). It remains unclear why we see this statistical difference between results obtained by forward encoding models (which predict neural data, given model states) and reverse (which predict model states, given neural data). Understanding this is a topic of ongoing investigation. Although these results are thus equivocal as to whether the agent and patient variables are represented in distinct anatomical locations, the underlying representational point remains: the present results provide further evidence that lmSTC represents abstract noun-role combinations (e.g., woman-as-agent and dog-as-patient) that are reused across sentences, enabling one to predict neural responses to new sentences based on previously observed responses to their parts.

Though these data provide evidence that lmSTC is encoding aspects of the event structure communicated by the sentence, the current results are agnostic as to what specific aspects of event structure may drive this organization. One possibility is that this organization reflects the underlying causal structure of the event. One might think that both the agent and the stimulus are, in some sense, the causal force behind what happens, while the patient and experiencer are the participants affected by what the agent/stimulus does. It is less obvious why this analysis should work for the experiencer-subject verbs we use, ('noticed' and 'detected'), however. These verbs are somewhat ambiguous with respect to who is responsible for originating the event. There may, however, be a more a general notion of causation that is applicable: one in which there is some asymmetric dependence of the relationship described between the two participants in the event. For example, take 'the moose noticed the hawk'. There would have been nothing for the moose to notice, had the hawk not had some pre-existing noteworthy feature. A related possibility is that this organization reflects the temporal, but not the causal, structure of the event described. In all the sentences used here, the event-relevant state of the agent and stimulus temporally precedes their entering into that relationship with the patient/experiencer. This could explain the generalization success on verbs such as 'noticed' that are not clearly causal. A third possibility is that these reflect the motion relations in the events. Movement, of course, is not a part of the core meaning of the verbs 'noticed' and 'surprised'. One can notice a typo, and be surprised by a scientific result. However, when the participants are mammals and birds, the most natural interpretation of the events involves the movement of the stimulus. This idea finds a theoretical basis in Jackendoff's (1992) theory of semantic structures. Jackendoff decomposes verb meanings into a set of primitive argument-taking semantic functions such as [GO(thing, place)] that recur across events. This makes it plausible that [GO(thing, place)] could be reused in interpreting the psych verbs here as well, even if one would not consider the entity's movement to be conveyed as part of the core meaning of the verb. On this view, the medial agent/stimulus region may represent the first argument of the GO function, the mover, and the patient/experiencer region might be part of the second argument of the GO function, which is the 'place' to which the entity moves. Finally, it is possible that these groupings reflect some general asymmetry in focus or salience that is related to, but distinct from syntactic structure (Baker 1997; Tversky 1977). That is, the patient/experiencer is the salient entity in the event and the agent/stimulus is simply defined with respect to that focal point. Adjudicating among these possibilities is beyond the scope of the present work. Future work thus requires careful attention to models of event structure (Jackendoff 1992; Levin and Hovav 2005; see Kemmerer et al. 2008 for an early model that predicts neural activity as a function of event-components).

Notably, we also found no evidence of semantic or phonological similarity structure of nouns within these roles in lmSTC. That is, the activity pattern for "moose" is no more similar to "cow" than it is to "goose," or vice versa. Although these null results must be interpreted with caution, they also suggest an intriguing third possibility: that lmSTC may not use similarity-based coding schemes, but instead uses an effectively arbitrary coding scheme. An arbitrary coding scheme maximizes symbol distinctiveness, and is a sign of an efficient, compressed code (Hopfield 1982). Characterizing such a code would be an important topic for future work.

We note that recent work on the neural basis of sentence processing has supported the theoretical integration of the lexicon and syntax (e.g., Fedorenko et al. 2011; Matchin and Hickok 2019), finding little separation between lexical and combinatorial operations in the brain. Our results may appear prima facie to be in opposition to this idea, given that we find that structure-dependent encoding models, and not bag-of-nouns models, predict BOLD signal in the regions we study. However, we take our results to be agnostic on this issue. The current evidence suggests that the regions we study here represent aspects of the underlying event structure, rather than sentence syntax. For example, we find that lmSTC activity is better explained by event structure than syntactic roles (see Fig. 4). Moreover, the body of prior work on mPFC (see preceding section) and the hippocampus is more consistent with the flexible encoding of conceptual representations of events that are derived from linguistic input, than with the lexico-syntactic operations constituting the derivational process.

"Repulsion" of Similar Event Representations in the Hippocampus

Given prior work implicating the hippocampus in relational representation (Cohen and Eichenbaum 1993; Davachi 2006), along with specific hypotheses (Duff and Brown-Schmidt 2012) and recent evidence (Piai et al. 2016; Blank et al. 2016) regarding its involvement in language comprehension tasks, we applied the same encoding models to an anatomically defined a priori hippocampal-ROI. Unlike the cortical regions, which we found to re-use structured representations across sentences, a subregion of left hippocampus tended to treat sentences containing the same parts as dissimilar, evidenced by below-chance generalization performance. The relevant "parts" are narrow, verb-specific conjunctions, thus exhibiting the opposite pattern as amPFC. Here, the response to sentences sharing a noun-verb conjunction (e.g., "the hawk surprised the moose" and "the hawk surprised the crow") are more dissimilar to one another than they are to other sentences that do not contain that conjunction. Although this effect is somewhat weak and should thus be interpreted with some caution, we believe it is credible for two reasons. First, it is supported by our beta-series connectivity analysis in which the hippocampal ROI exhibits trial-by-trial anticorrelation with generalization performance in amPFC: the better the model predicts activity to a novel sentence in amPFC on a particular trial, the worse it generalizes in the hippocampus. This effect is strong ($P = 2.37 \times 10^{-7}$). Second, it is strikingly consistent with an emerging body of work documenting other "repulsive" effects in the hippocampus (Schapiro et al. 2012; Schlichting et al. 2015; Favila et al. 2016; Chanales et al. 2017), in which similar states come to be mapped to dissimilar (rather than simply orthogonal) encodings. Broadly, pattern separation decreases the representational overlap between two states within a downstream neural population (such as hippocampus), when compared to the overlap of those states in the upstream region (here, entorhinal cortex). It is most well-characterized in the dentate gyrus (DG) and CA3 subfields of the hippocampus. Thus, two DG representations (state 1 and state 2) are expected to be less correlated than the corresponding entorhinal cortex representations that produce them. This functionality is hypothesized to improve the likelihood of successful memory recall (Marr 1969; Treves and Rolls 1992; O'Reilly and McClelland 1994). However, in its canonical formulation, pattern separation entails only the orthogonalization of codes, achieved using a sparse, high-dimensional representation in DG/CA3. Our work adds to a growing body of recent fMRI studies that suggest that, at least when measured at the voxel level, closely relates states are sometimes not simply orthogonalized in the hippocampus, but are anticorrelated in certain environments (e.g., Favila et al. 2016; Chanales et al. 2017).

For example, Favila et al. (2016) find evidence that images paired during a learning phase are later encoded as more dissimilar to one another than nonpaired images. They find that this separation of similar events (there, scene and face images) promotes associative learning by reducing overlap-driven interference, consistent with the hypothesized function of pattern separation. Chanales et al. (2017) report complementary effects in a spatial navigation domain. There, hippocampal representations of overlapping routes through an environment become more dissimilar to one another than nonoverlapping routes over the course of learning. Re-use of a route component thus has a "repulsive" effect on the hippocampal encoding (but not the encoding in other task-relevant regions) (Chanales et al. 2017). In the present work, we find evidence that sentences sharing noun-verb conjunctions tend to be encoded more dissimilarly than sentences that do not share these conjunctions in an anterior region of left hippocampus. Taken in combination with our findings regarding amPFC, this general pattern is consistent with the recent idea that parts of mPFC enable rapid traversals of conceptual, as well as physical space (Behrens et al. 2018).

At a high level, it may seem unlikely that representations of similar states would actually be anticorrelated (rather than simply noncorrelated). However, it is noteworthy that other repulsive processes are known in the physical and biological sciences, such as the distribution of fermions at thermal equilibrium (Kulesza and Taskar 2012), the distinctive phenotypes of species of Eurasian nuthatches occupying the same geographic region (Brown and Wilson 1956), (which may be thought of as repulsion in trait-space), and the spatial distributions of neighboring termite mounds (Martin et al.

2018). Computationally, these phenomena may all be subsumed under a class of probabilistic models known as "determinantal point processes" (DPPs) (Kulesza and Taskar 2012). In DPP models, the more similar two feature vectors (thus having a small determinant) the less likely they are to succeed/neighbor one another in time/space. Here, the relevant "space" is the hippocampal code space, in which potentially similar encodings are repelled. We suggest that DPPs may be useful high-level models of the repulsive effects in hippocampal codes, though this remains a topic for future work.

Our initial motivation for considering the hippocampus was its ability to rapidly form high-dimensional conjunctive representations that encode relations between different combinations of inputs (Cohen and Eichenbaum 1993; Treves and Rolls 1992; O'Reilly and McClelland 1994). However, it is unclear from the present results what role the hippocampus plays in naturalistic language processing. Should we expect the same pattern separation signature to occur in naturalistic contexts that lack the strong semantic similarity between successive sentences employed here? It seems possible that the hippocampal effects we observe reflect the discrimination of highly similar sentence meanings (event representations) in close succession, caused by the menu-like structure of the current experiment (see Fig. 2). Relevant evidence also comes from Duff et al. (2011), who had patients with hippocampal amnesia generate distinct linguistic labels for novel shapes. Notably, patients with hippocampal amnesia were impaired relative to controls in labeling similar, but not dissimilar shapes, suggesting that hippocampus may be recruited for linguistic contexts when the separation of similar inputs is driven by the similarity structure of the sentences over time.

However, although the effect may owe to the similarity structure of stimulus space we employ, we note that it does not appear to depend on the explicit task subjects performed in the scanner. As with previous empirical demonstrations of hippocampal pattern separation-like phenomena (Bakker et al. 2008; Schapiro et al. 2012; Favila et al. 2016) we note that this particular representation (here, narrow roles) is not directly tied to subjects' explicit task. The task required the extraction and maintenance of more abstract role information (who did it?/to whom was it done?), but did not require maintenance of verb-specific information. In this, it seems to be driven by the overall similarity of particular aspects of the content, rather than by the similarity of the response. Note also that the lmSTC effects do not appear to be task-dependent. A task-dependent account would predict that the anatomical separation should be better modeled by grouping roles to form subject/object categories (what the task queried) rather than agent (stimulus)/patient (experiencer) categories, which sometime cross subject/object categories (e.g., "the moose surprised the crow" (in which the experiencer is the object) and "the crow noticed the crow" (in which the experiencer is the subject)).

The hippocampal effect here falls within the anterior portion of the a priori anatomical ROI. This location may seem at odds with suggestions that "posterior" hippocampus is involved in separating representations, while anterior hippocampus supports generalization across experiences (e.g., Collin et al. 2015; Schlichting et al. 2015). We briefly consider two possible (related) reasons why we may see the current effect in anterior, but not posterior hippocampus. First, we note that other observations of pattern separation signatures in anterior (as well as posterior) hippocampus have involved relatively weak statistical regularities between the associated pairs (Schapiro et al. 2012) using, for example, inter-mixed, rather than block, learning paradigms (Schlichting et al. 2015). This is analogous to the current regime in which particular conjunctions (e.g., hawk-noticed) are relatively infrequent in the experiment and randomly presented. Moreover, outside of the experimental context, the relevant conjunctive representations are semantically weak (given that there is a noticing event, the probability that the entity doing the noticing is a hawk (P(entity-type | action-type), or the probability that the thing a hawk does is notice (P(action-type | entity-type) will be quite low). Though this may partly explain why we do see such effects in anterior hippocampus, it does not explain the lack of an effect in posterior hippocampus. We speculate that this may be due to an additional anatomical constraint in which anterior, but not posterior hippocampus is responsive to particular types of representational structures. For example, Blank et al. (2016) find that anterior, but not posterior hippocampus is implicated in univariate contrasts of sentence-level linguistic processing. It is intriguing that anterior hippocampus is also involved in the representation of other highly structured forms, such as social hierarchies (Kumaran et al. 2012, 2016). Anterior hippocampus thus appears to play a role in weakly associated and perhaps also richly structured domains, such as sentence processing. This remains speculative, however, and an important topic for future work.

Conclusion

By contrasting the performance of encoding models operating at different levels of abstraction, we provide evidence that the brain employs complementary strategies for encoding who did what to whom. A region of amPFC encodes narrow verb-specific conjunctions (woman-as-chaser), reused across sentences. This differs from a region of lmSTC, which carries information about broad roles ("agent—'the woman did something'"; "patient—'the dog had something done to it'"). The success of different encoding models in subregions of the lmSTC and amPFC may reflect a trade-off between abstraction (lmSTC) and specificity (amPFC) in the deployment of reusable representations of event structure. Broad roles could support generalization to novel verbs and the mapping of event structure to sentence syntax. Narrow roles, by contrast, may provide structured semantic pieces necessary to imagine and reason about more specific events. We thus interpret our effects as two different ways to build a thought: one uses abstract low-dimensional role representations, invariant across classes of verbs and supported by the lateral temporal lobe. The other extracts specific subcomponents (here, the meanings of verb-noun combinations) that recur across contexts, perhaps using statistical learning. Critically, both strategies use and reuse familiar parts according to combinatorial rules ("who did the chasing? who was chased?").

It is notable that the effects observed in the amPFC reflect a combination of representational strategies traditionally associated with classical symbolic systems, on the one hand, and feedforward neural networks, on the other (cf. Pinker 1997; Marcus 2001). Classical systems have historically favored abstract representations and dynamic variable binding. For example, a mathematical formula allows for the binding of arbitrary values to variables (but see Doumas et al. 2008; Kriete et al. 2013; Graves et al. 2016 for network models of binding). Feedforward neural networks, by contrast, typically store and retrieve specific conjunctive representations, for example, conjoining multiple edges in one layer to form a contour one layer up. In the amPFC, the representations are conjunctive, representing "cow-as-approacher," distinct from the simultaneous representation of "cow" and "approached." And yet these conjunctive representations must be dynamically bound to other semantic elements, such that the same conjunctive representation is reused in "cow approached crow" and "hawk was approached by cow." This suggests an intriguing possibility: that the representations in amPFC function like bits of conceptual "clip art," hybrid units that can be mixed and matched like symbols, but that also encode conceptual content reflective of conjunction-specific features.

Though we suggest that amPFC and lmSTC reflect two different ways to represent event relations, a number of qualifications are in order. First, we do not mean to suggest that these are the only ways that the brain might encode relations between event participants, or that this is an exhaustive study of either the types of relations (i.e., event-types) or the entities (i.e., a small set of mammals and birds) that they hold between. Nor do we mean to suggest that the particular regions that we study constitute a complete list of those involved in mapping from syntactically structured input to a nonlinguistic event representation. Specifically, the inferior prefrontal cortex (Hagoort et al. 2004), middle temporal gyrus (Dronkers et al. 2004) and angular gyrus (Boylan et al. 2015; Williams et al. 2017,) are particularly likely to support aspects of this mapping. More generally, our claims are about the existence of what we have observed in different brain regions, not about the uniqueness of what we have observed.

Finally, these results do not speak to whether these regions themselves implement flexible binding mechanisms, able to generate novel role-filler bindings on the fly (see Smolensky 1990; Plate 1995; Hummel and Holyoak 2003; Doumas et al. 2008; Kriete et al. 2013), or whether they reflect conceptual combinations that are computationally bound elsewhere, or simply retrieved from memory. The particular methods we employ here target the nature of the representation, not the process that creates it. Here, we show that two regions (amPFC and lmSTC) are involved in representing who did what to whom in such a way that these role-dependent representations are reused across sentences and differ in their abstraction. Understanding how the brain adaptively coordinates these representational systems to produce a unified understanding of novel, complex events remains an important goal for future research.

Supplementary Material

Supplementary material is available at *Cerebral Cortex* online

References

Anderson AJ, Binder JR, Fernandino L, Humphries CJ, Conant LL, Aguilar M *et al.* 2016. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cereb Cortex*. 27:4379–4395.

Anderson AJ, Lalor EC, Lin F, Binder JR, Fernandino L, Humphries CJ, Wang X. 2018. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cereb Cortex*.

Anzellotti S, Coutanche MN. 2018. Beyond functional connectivity: investigating networks of multivariate representations. *Trends in cognitive sciences*. 22:258–269.

Bakker A, Kirwan CB, Miller M, Stark CE. 2008. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science*. 319:1640–1642.

Baker MC. 1997. Thematic roles and syntactic structure. In: *Elements of grammar*. Dordrecht: Springer, pp. 73–137.

Barron HC, Dolan RJ, Behrens TE. 2013. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat Neurosci*. 16:1492–1498.

Bedny M, Caramazza A, Grossman E, Pascual-Leone A, Saxe R. 2008. Concepts are more than percepts: the case of action verbs. *J Neurosci*. 28:11347–11353.

Behrens TE, Muller TH, Whittington JC, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z. 2018. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*. 100:490–509.

Belletti A, Rizzi L. 1988. Psych-verbs and θ-theory. *Nat Lang Linguist Theory*. 6:291–352.

Belin P, Zatorre RJ, Ahad P. 2002. Human temporal-lobe response to vocal sounds. *Cognit Brain Res*. 13:17–26.

Bemis DK, Pylkkänen L. 2011. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J Neurosci*. 31:2801–2814.

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

Blank I, Duff MC, Brown-Schmidt S, Fedorenko E. 2016. Expanding the language network: domain-specific hippocampal recruitment during high-level linguistic processing. *bioRxiv*. doi: 10.1101/091900.

Bowman CR, Zeithamova D. 2018. Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *J Neurosci*. 2811–2817.

Boylan C, Trueswell JC, Thompson-Schill SL. 2015. Compositionality and the angular gyrus: a multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia*. 78:130–141.

Brown WL, Wilson EO. 1956. Character displacement. *Systematic zoology*. 5:49–64.

Bunge SA, Helskog EH, Wendelken C. 2009. Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *Neuroimage*. 46:338–342.

Chanales AJ, Oza A, Favila SE, Kuhl BA. 2017. Overlap among spatial memories triggers repulsion of hippocampal representations. *Curr Biol*. 27:2307–2317.

Chao LL, Haxby JV, Martin A. 1999. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*. 2:913.

Cohen NJ, Eichenbaum H. 1993. *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.

Collin SH, Milivojevic B, Doeller CF. 2015. Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience*. 18:1562.

Coutanche MN, Thompson-Schill SL. 2013. Informational connectivity: identifying synchronized discriminability of multivoxel patterns across the brain. *Front Hum Neurosci*. 7:15.

Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*. 29:162–173.

Davachi L. 2006. Item, context and relational episodic encoding in humans. *Curr Opin Neurobiol*. 16:693–700.

Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C. 2015. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*. 88:2–19.

Doumas LA, Hummel JE, Sandhofer CM. 2008. A theory of the discovery and predication of relational concepts. *Psychol Rev*. 115:1–43.

Dowty D. 1991. Thematic proto-roles and argument selection. *Language*. 67:547–619.

Dronkers NF, Wilkins DP, Jr VV, RD Redfern BB, Jaeger JJ. 2004. Lesion analysis of the brain areas involved in language comprehension. *Cognition*. 92:145–177.

Duff MC, Warren DE, Gupta R, Vidal JP, Tranel D, Cohen NJ. 2011. Teasing apart tangrams: testing hippocampal pattern separation with a collaborative referencing paradigm. *Hippocampus*. 22:1087–1091.

Duff MC, Brown-Schmidt S. 2012. The hippocampus and the flexible use and processing of language. *Front Hum Neurosci*. 6:69–80.

Eichenbaum H. 1999. The hippocampus and mechanisms of declarative memory. *Behav Brain Res*. 103:123–133.

Elli GV, Lane C, Bedny M. 2019. A double dissociation in sensitivity to verb and noun semantics across cortical networks. *Cereb Cortex*. 1–15.

Favila SE, Chanales AJ, Kuhl BA. 2016. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun*. 7:11066.

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

Fedorenko E, Behr MK, Kanwisher N. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proc Natl Acad Sci*. 108:16428–16433.

Fodor JA, Pylyshyn ZW. 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition*. 28:3–71.

Fillmore CJ. 1967. *The case for case*.

Frege G, Patzig G. 2003. *Logische untersuchungen*. Vol 4031. Vandenhoeck & Ruprecht.

Frankland SM, Greene JD. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proc Natl Acad Sci*. 112:11732–11737.

Frankland SM, Greene JD. 2019. Concepts and compositionality: in search of the brain's language of thought. *Annu Rev Psychol*. doi: 10.1146/annurev-psych-122216-011829.

Gertner Y, Fisher C, Eisengart J. 2006. Learning words and rules: abstract knowledge of word order in early sentence comprehension. *Psychol Sci*. 17:684–691.

Goldberg AE. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Graves WW, Binder JR, Desai RH, Conant LL, Seidenberg MS. 2010. Neural correlates of implicit and explicit combinatorial semantic processing. *Neuroimage*. 53:638–646.

Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J *et al*. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*. 7626: 471–476.

Green AE, Kraemer DJ, Fugelsang JA, Gray JR, Dunbar KN. 2009. Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cereb Cortex*. 20:70–76.

Hagoort P, Hald L, Bastiaansen M, Petersson KM. 2004. Integration of word meaning and world knowledge in language comprehension. *Science*. 304:438–441.

Hannula DE, Ranganath C. 2008. Medial temporal lobe activity predicts successful relational memory binding. *J Neurosci*. 28:116–124.

Hartshorne JK, O'Donnell TJ, Sudo Y, Uruwashi M, Lee M, Snedeker J. 2016. Psych verbs, the linking problem, and the acquisition of language. *Cognition*. 157:268–288.

Hopfield JJ. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci*. 79:2554–2558.

Hummel JE, Holyoak KJ. 2003. A symbolic-connectionist theory of relational inference and generalization. *Psychol Rev*. 10:220.

Humphries C, Binder JR, Medler DA, Liebenthal E. 2006. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J Cogn Neurosci*. 18:665–679.

Huth AG, Nishimoto S, Vu AT, Gallant JL. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*. 76:1210–1224.

Huth AG, de WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*. 532:453.

Jackendoff R. 1992. *Semantic structures*. Cambridge, MA: MIT press.

Kemmerer D, Castillo JG, Talavage T, Patterson S, Wiley C. 2008. Neuroanatomical distribution of five semantic components of verbs: evidence from fMRI. *Brain Lang*. 107:16–43.

Knowlton BJ, Morrison RG, Hummel JE, Holyoak KJ. 2012. A neurocomputational system for relational reasoning. *Trends Cognit Sci*. 16:373–381.

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*. . doi: 10.1073/pnas.1303547110.

Kulesza A, Taskar B. 2012. Determinantal point processes for machine learning. *Found Trends Mach Learn*. 5:123–286.

Kumaran D, Summerfield JJ, Hassabis D, Maguire EA. 2009. Tracking the emergence of conceptual knowledge during human decision making. *Neuron*. 63:889–901.

Kumaran D, Melo HL, Duzel E. 2012. The emergence and representation of knowledge about social and nonsocial hierarchies. *Neuron*. 76:653–666.

Kumaran D, Banino A, Blundell C, Hassabis D, Dayan P. 2016. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron*. 92:1135–1147.

Levin B, Hovav MR. 2005. *Argument realization*. Cambridge, UK: Cambridge University Press.

Libby LA, Hannula DE, Ranganath C. 2014. Medial temporal lobe coding of item and spatial information during relational binding in working memory. *J Neurosci*. 34:14233–14242.

Martin SJ, Funch RR, Hanson PR, Yoo EH. 2018. A vast 4,000-year-old spatial pattern of termite mounds. *Curr Biology*. 28:R1292–R1293.

Mazoyer BM, Tzourio N, Frak V, Syrota A, Murayama N, Levrier O, Mehler J. 1993. The cortical representation of speech. *J Cogn Neurosci*. 5:467–479.

Marcus GF. 2001. *The algebraic mind: integrating connectionism and cognitive science*. Cambridge, MA: MIT press.

Marr D. 1969. A theory of cerebellar cortex. *J Physiol*. 202:437–470.

Matchin W, Hickok G. 2019. The cortical organization of syntax. *Cereb Cortex*.

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*. 5880:1191–1195.

Montague R. 1970. Universal grammar. *Theoria*. 36:373–398.

Mueller ST, Seymour TL, Kieras DE, Meyer DE. 2003. Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *J Exp Psychol Learn, Mem Cogn*. 29:1353–1380.

Mumford JA, Turner BO, Ashby FG, Poldrack RA. 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*. 59:2636–2643.

O'Reilly RC, McClelland JL. 1994. Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus*. 4:661–682.

Pallier C, Devauchelle AD, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci*. 108:2522–2527.

Peelen MV, Romagno D, Caramazza A. 2012. Independent representations of verbs and actions in left lateral temporal cortex. *J Cogn Neurosci*. 24:2096–2107.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N *et al*. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nat Commun*. 9:963–976.

Pesetsky D. 1987. Binding problems with experiencer verbs. *Linguistic Inquiry*. 18:126–140.

Piai V, Anderson KL, Lin JJ, Dewar C, Parvizi J, Dronkers NF, Knight RT. 2016. Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. *Proc Natl Acad Sci*. 113:11366–11371.

Pinker S. 1989. *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: MIT press.

Pinker S. 1997. *How the mind works*. New York, NY: Norton.

Plate TA. 1995. Holographic reduced representations. *IEEE Trans Neural Networks*. 6:623–641.

Poeppel D, Guillemin A, Thompson J, Fritz J, Bavelier D, Braun AR. 2004. Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. *Neuropsychologia*. 42:183–200.

Preston AR, Eichenbaum H. 2013. Interplay of hippocampus and prefrontal cortex in memory. *Curr Biol*. 23:764–773.

Price CJ, Moore CJ, Humphreys GW, Wise RJ. 1997. Segregating semantic from phonological processes during reading. *J Cogn Neurosci*. 9:727–733.

Pylkkänen L, McElree B. 2007. An MEG study of silent meaning. *J Cogn Neurosci*. 19:1905–1921.

Pylkkänen L. 2008. Mismatching meanings in brain and behavior. *Lang Linguist Compass*. 2:712–738.

Pylkkänen L. 2019. The neural basis of combinatory syntax and semantics. *Science*. 366:62–66.

Ramnani N, Owen AM. 2004. Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature reviews neuroscience* 5:184–94.

Ranganath C, D'Esposito M. 2001. Medial temporal lobe activity associated with active maintenance of novel information. *Neuron*. 31:865–873.

Rissman J, Gazzaley A, D'Esposito M. 2004. Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*. 23:752–763.

Rodd JM, Vitello S, Woollams AM, Adank P. 2015. Localising semantic and syntactic processing in spoken and written language comprehension: an activation likelihood estimation meta-analysis. *Brain Lang*. 141:89–102.

Schapiro AC, Kustner LV, Turk-Browne NB. 2012. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr Biol*. 22:1622–1627.

Schlichting ML, Mumford JA, Preston AR. 2015. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat Commun*. 6:8151–8161.

Schuler KK. 2005. *VerbNet: a broad-coverage, comprehensive verb lexicon*.

Selfridge OG. 1958. *Pandemonium: a paradigm for learning*.

Smolensky P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif Intell*. 46:159–216.

Thompson-Schill SL. 2003. Neuroimaging studies of semantic memory: inferring "how" from "where". *Neuropsychologia*. 41:280–292.

Tomasello M. 1992. *First verbs: a case study of early grammatical development*. Cambridge, UK: Cambridge University Press.

Treves A, Rolls ET. 1992. Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*. 2:189–199.

Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER *et al*. 2007. Schemas and memory consolidation. *Science*. 316:76–82.

Tse D, Takeuchi T, Kakeyama M, Kajii Y, Okuno H, Tohyama C *et al*. 2011. Schema-dependent gene activation and memory encoding in neocortex. *Science*. 333:891–895.

Tversky A. 1977. Features of similarity. *Psychol Rev*. 84:327.

Urbanski M, Bréchemier ML, Garcin B, Bendetowicz D, Thiebaut de Schotten M, Foulon C *et al*. 2016. Reasoning by analogy requires the left frontal pole: lesion-deficit mapping and clinical implications. *Brain*. 139:1783–1799.

Van Valin RD Jr, Van Valin RD. 2005. *Exploring the syntax-semantics interface*. Cambridge, UK: Cambridge University Press.

Vandenberghe R, Nobre A, Price C. 2002. The response of left temporal cortex to sentences. *J Cogn Neurosci*. 14:550–560.

Vigneau M, Beaucousin V, Herve PY, Duffau H, Crivello F, Houde O *et al*. 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*. 30:1414–1432.

Volle E, Gilbert SJ, Benoit RG, Burgess PW. 2010. Specialization of the rostral prefrontal cortex for distinct analogy processes. *Cereb Cortex*. 20:2647–2659.

Wang J, Cherkassky VL, Yang Y, Chang KMK, Vargas R, Diana N, Just MA. 2016. Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cogn Neuropsychol*. 33:257–264.

Wang J, Cherkassky VL, Just MA. 2017. Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states. *Hum Brain Mapp*. 38:4865–4881.

Williams A, Reddigari S, Pylkkänen L. 2017. Early sensitivity of left perisylvian cortex to relationality in nouns and verbs. *Neuropsychologia*. 100:131–143.

Wu DH, Waller S, Chatterjee A. 2007. The functional neuroanatomy of thematic role and locative relational knowledge. *J Cogn Neurosci*. 19:1542–1555.

Just MA, Wang J, Cherkassky VL. 2017. Neural representations of the concepts in simple sentences: concept activation prediction and context effects. *Neuroimage*. 157: 511–520.

Yang Y, Wang J, Bailer C, Cherkassky V, Just MA. 2017. Commonality of neural representations of sentences across languages: predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. *Neuroimage*. 146:658–666.

Ziegler J, Snedeker J. 2018. How broad are thematic roles? Evidence from structural priming. *Cognition*. 179:221–240.

Zeithamova D, Dominick AL, Preston AR. 2012. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*. 75: 168–179.