Man Bites Dog: The Representation of Structured Meaning in Left-Mid Superior Temporal Cortex

Man bites dog: The representation of structured meaning in left-mid superior temporal cortex

A dissertation presented

by

Steven Michael Frankland

To the Department of Psychology

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

In the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May 2015

Dissertation Advisor: Professor Joshua D. Greene Steven Michael Frankland

Man bites dog: The representation of structured meaning in left-mid superior temporal cortex

Abstract

Human brains flexibly combine the meanings of individual words to compose structured thoughts. For example, by combining the meanings of ‘bite’, ‘dog’, and ‘man’, we can think either of a dog biting a man, or the newsworthy case of a man biting a dog (Pinker, 1997). Here, in three functional Magnetic Resonance Imaging (fMRI) experiments, we identify a region of left-mid Superior Temporal Cortex (lmSTC) that represents the current values of abstract semantic variables (“Who did it?” and “To whom was it done?”) in anatomically distinct sub-regions. Experiment 1 first identifies a broad region of lmSTC whose activity patterns (a) facilitate decoding of who did what to whom and (b) predict affective amygdala responses that depend on this information (e.g. “the baby kicked the grandfather” vs. “the grandfather kicked the baby”). Experiment 2 then identifies distinct, but neighboring, sub-regions of lmSTC whose activity patterns carry information about the identity of the current *agent* (“Who did it?”) and the current *patient* (“To whom was it done?”). These neighboring sub-regions lie along the upper bank of the superior temporal sulcus and the lateral bank of the superior temporal gyrus, respectively. At a high-level, these regions may function like topographically defined data registers, encoding the fluctuating values of abstract semantic variables. Experiment 3 replicates the agent/patient topography of Experiment 2, and further suggests that these variables do not represent the grammatical relations of the sentence, but the semantic relations of the participants in the event described. The code by which lmSTC encodes the values of these variables remains unclear, however. We find no positive evidence that it is either phonological or semantic, leaving open the possibility that lmSTC prioritizes distinctiveness and efficiency by using a compressed code. This functional architecture, which in key respects resembles that of a classical computer, may play a critical role in enabling humans to flexibly generate complex thoughts.

Table of Contents

Acknowledgements

Thanks, first, to my advisor Josh Greene. I'm deeply grateful for Josh's willingness to work with me on these projects. I hope that some of his ability to identify, eloquently articulate, and courageously pursue interesting research has rubbed off. More generally, Josh has been as decent, compassionate, and supportive an advisor as one could ask for.

I am also indebted to Susan Carey. Susan is unparalleled in the generosity of time and intellect she provides students, the department, and the field. I've been fortunate to learn from her. Fiery Cushman has offered characteristically insightful feedback at many points throughout this process. His thoughtful comments have influenced both the conceptualization of these problems and the specifics of our approach. Steve Pinker has provided a wealth of encouragement and rich intellectual context for this research. Steve has done as much as anyone to bring the issues motivating this thesis in to focus, and his intellectual influence extends well beyond the title. Alfonso Caramazza has provided incisive and actionable advice, and our analyses are more thorough as a result.

Thanks to Anita Murrell, Sarah Coughlon, Xi Yang, and Rebecca Fine for research assistance, and to all the members of the Greene and Cushman labs who helped with scanning. Thanks also to the 100+ subjects who participated in this research. It's no fun to spend 90 minutes reading about hogs approaching hawks while lying immobilized in a magnet. I appreciate your help.

Thanks to the Greene lab members past and present for providing feedback and, perhaps more importantly, friendship throughout graduate school. Special thanks to Regan Bernhard, Donal Cahill, Alek Chakroff, Joe Paxton, Ann Carroll, Sara Gottlieb, Morgan Henry, and Elinor Amit. I couldn't have asked for better lab mates. Thanks also to Roman Feiman, Anna Leshinskaya, Jonathan Phillips, and Jorie Koster-Hale for valuable discussions.

Thanks to the National Science Foundation, and the Harvard Mind, Brain, and Behavior initiative for generous funding.

I owe an inexpressible amount to my parents, Mike and Chris Frankland. Their love and support has enabled me to pursue my interests, however impractical they may be.

Finally, my deepest thanks to my wife, Patience Gallagher. Patience has served many of the roles mentioned above; from advisor and critic, to research assistant and scan subject. Her most (in)valuable contribution has been, and continues to be, her love.

Yesterday the world's tallest woman was serenaded by thirty pink elephants. The previous sentence is false, but perfectly comprehensible, despite the improbability of the situation it describes. It's comprehensible because the human mind can flexibly combine the meanings of individual words ("woman", "serenade", "elephants", etc.) to compose structured thoughts, such as the meaning of the aforementioned sentence. Our minds make sense of such complex expressions by adhering to an elegant principle: the meaning of a complex expression is a function of the meaning of its parts and the way in which those parts are combined. This is the principle of compositionality (Frege, 1892/1977). It is what allows us to entertain new and complex thoughts. But how? The psychology of compositionality introduces rich questions about the mind and its realization in the brain: What are the basic components of our minds, such that we can draw ideas from a reusable stock and nimbly assemble them to form new semantic structures? And how does the brain, equipped with a finite number of neurons, implement these combinatorial operations to encode the vast number of thinkable thoughts?

But while these issues are theoretically weighty, the empirical evidence is light, particularly concerning the brain's strategies for building complex thoughts. The present thesis consists of three functional neuroimaging studies that address how the brain flexibly encodes the meaning of simple sentences involving an agent ("who did it?"), an action ("what was done?"), and a patient ("to whom was it done?"). Before describing this empirical work, we review the relevance of compositionality to our understanding of the mind's architecture, as well as the various theoretical models for encoding complex, structured meanings in computational systems.

Compositionality and Cognitive Architecture

Thinking is a mechanical process. But what are the basic parts of our mental machinery? Important clues come from patterns in the thoughts we can entertain (Fodor & Pylyshyn, 1988; Pinker, 1997; Fodor, 1998). Take, for example, the sentence “Drew Gilpin Faust beat Larry Summers in an arm wrestling match”. Anyone who can understand the previous sentence can also understand “Larry Summers beat Drew Gilpin Faust in an arm wrestling match”. If that mind also has a concept of *Joe Biden*, then it will also understand “Joe Biden beat Larry Summers in an arm wrestling match”, and, probably even “Joe Biden beat a horse in a foot race”, if it also has the concepts *horse* and *foot race.*

Why do we not find minds that can understand “Faust defeated Summers”, but not “Summers defeated Faust”? The classical view of the mind’s architecture has a simple and compelling answer. On the classical view, human intelligence consists in manipulating the symbols in an inner language, or perhaps, languages (Newell, 1980; Fodor, 1975; Pinker, 1997). These symbols represent aspects of the actual or possible world, and can be combined and recombined to form new, often more complex, symbol-structures. The reason anyone who can understand “Faust defeated Summers”, can also understand “Summers defeated Faust” is that the meanings of both sentences are mentally represented by assigning the same reusable symbols (*Faust, Summers*) to the same reusable argument positions (“who did the defeating?”/ “whom was defeated?”); just with the mapping between symbols and positions reversed. If you have a stock of concepts and combinatorial mechanisms for assembling them, you can churn out a vast number of structured thoughts. Composing structured representations using the classicist inventory of parts is thus straightforward. Indeed, Fodor & Pylyshyn (1988) take the possession of complex representations with combinatorial structure to be partly definitional of classical cognitive architecture. Part of what makes the classical view of the mind the classical view just is that there are systematic procedures for combining and recombining symbols in a mental language [1] .

This is not the only possible conception of the mind's essential components, however. An alternative view, known as connectionism, emerged in the 1980s amongst neural network modelers (Rumelhart & McClelland, 1988). On this view, the basic parts of the mind were not symbols and explicit procedures for manipulating them, but interconnected nodes that can vary in activity level, modifiable connections between these nodes, and a collection of learning algorithms that tweak the strength of those connections to make the network smarter. In the hands of some researchers, these networks were not merely intended to model how the brain implements classical architectures in neuron-like networks; they were meant as an alternative conception of the computational underpinnings of intelligence (Churchland, 1996; Clark, 1993). The essence of intelligence, on the connectionist view, is the mind's ability to learn statistical regularities and contingencies from experience, and apply this learning to new contexts.

Neural network models have a number of desirable properties. They learn from experience. They naturally treat similar things as similar, and generalize accordingly. And, if you break them, they degrade in ways that sometimes mirror the cognitive degradation accompanying neurological damage (Rogers & McClelland, 2004). These virtues do not, however, necessarily make them adequate as a general cognitive architecture. They have

[1] The second component of Fodor & Pylyshyn's (1988) definition of classical architecture is a commitment to structure-sensitive cognitive processes. These are operations that depend upon the form of a representation, such as various types of logical inference (P&Q→P).

trouble explaining a few basic cognitive functions, including an inability to model the English past tense (Pinker & Prince, 1988;), and an inability to represent individuals as such, distinct from categories (Pinker, 1997; Marcus, 2001). Prominent among these limitations, however, is their inability to explain the compositionality of thought and language (Fodor & Pylyshyn, 1988; Fodor & McLaughlin, 1988; Pinker, 1997; Marcus, 2001; Hummel et al. 2000). In the simplest neural network models, called 'perceptrons', the individual nodes in the network's output layer of node represent features that can be gleaned from the input. These may be concepts of who the participants in the event were (Faust, Summers), or and what type of event it was (arm wrestling contest). But, of course, simply activating the concept-nodes for 'Faust', 'defeated' and 'Summers' in response to an event does not tell you who was the victor, and who was the vanquished. Perhaps, one might think, the representational nodes could be pre-specified to represent higher-order variable/value pairs (Faust-as-defeater), or even entire propositions (defeat(Faust, Summers)). However, as we will see below, such models have difficulty explaining how we can understand new combinations, and why the thoughts we can entertain cluster systematically.

Since these early realizations, modelers have found clever and productive ways to implement structured representations in neural networks (Smolensky, 1990; Ajjanagadde & Shastri, 1991; Plate, 1995, Hummel & Holyoak, 1997; 2003, O'Reilly & Busby, 2002; and most recently Kriete et al., 2013). These models vary in their strengths and limitations, which we will discuss below, but to the extent that they succeed, it is as implementations of classical architecture, not as alternatives (Pinker & Prince, 1988; Marcus, 1998). A mental toolbox of representational nodes, connection strengths, and learning rules doesn't provide an alternative means through which to systematically encode structured information, such as who did what to whom in an event, without somehow implementing symbol structures. This point is increasingly recognized in motivating theoretical models (Hummel et al., 2000; Doumas et al., 2008; Kriete et al., 2013), and debates about compositionality and cognitive architecture have quieted (Fodor & McLaughlin, 1990. Smolensky, 1991; Fodor, 1997).

Theorists interested in modeling the mental and neural bases of compositionality are now, instead, largely concerned with one key question: how are simpler representations glued together to form more complex ones? (Hummel & Holyoak, 1997; 2003; Hummel et al., 2000; O'Reilly & Busby, 2002; Doumas et al., 2008; Hayworth, 2012; Van der Velde & de Kamps, 2007; Kriete, et al. 2013; Haysworth, 2012; Stewart et al., 2011). This question, which has its roots in the cognitive neuroscience of vision, is often referred to as "the binding problem". Writers refer to a number of related problems as "*the* binding problem", (Shadlen & Movshon, 1999). However, at root, these problems all concern how neurons flexibly integrate the information carried by distinct neural populations to encode a single complex stimulus. It is a problem because it's implausible to think that the brain devotes individual neurons, sometimes called "cardinal neurons" (Martin, 1994; Shadlen & Movshon, 1999), to each and every combination of stimulus features we could encounter. This binding problem recurs in many contexts throughout cognitive neuroscience. For example, how does the brain signal that different visual features (color, shape, motion, etc.) belong to one and the same object, despite the fact that the feature-information is dispersed throughout visual cortex? Likewise, how are the various features of a conceptual representation (shape, function, etc.), which are likely housed in different brain areas

(Patterson, Nestor, & Rogers, 2007), coordinated to represent one concept? And, of most relevance here, how are representations of semantic variables and their values (often called "roles" and "fillers") bound to one another, and so as to represent a single proposition? These neural binding problems arise whenever independent neural representations need to be mixed and matched to form different complex representations.

It is assumed that the problem of how the brain encodes who did what to whom is identical, in key respects, to the other binding problems, for example, in vision (Hummel et al., 2000; Doumas et al., 2008; Stewart & Eliasmith, 2012; Van der Velde & de Kamps, 2006). Notice, however, that the problem of role/filler binding is characterized abstractly in terms of representational content, rather than in terms of known neural representations. This is because we simply don't know how information about semantic roles is encoded neurally. We can't say: "how is information about role, coded in the pineal gland, bound to the symbolic representations distributed throughout cortex?" It may be the case that role information is encoded in one part of the brain, and filler information elsewhere. But, despite widespread assumption (Hummel et al., 2000; Doumas et al., 2008; Stewart & Eliasmith, 2008; Van der Velde, 2006), we don't know. Below, we review the main theoretical models that have been proposed for encoding complex, structured meanings in the brain, before presenting our empirical work targeting the neural representation of these meanings.

Theoretical Models

Conjunctive Coding

We noted above that neural network models that only contain nodes for individual concepts, without respect to the roles those concepts plays in a sentence or event, cannot represent who did what to whom. The most direct way to extend this model is to add individual nodes, or sets of nodes, that selectively encode the presence of concepts in particular roles. This strategy is known as conjunctive coding. To encode "the man bit", the node or nodes for "man as biter" would be activated when and only when 'man' is present, AND is the one in the event doing the biting. Such models might then attempt to encode an entire proposition (e.g., "the man bit the dog") using the simultaneous activation of the nodes encoding "man as biter" and "dog as bitee". Or, perhaps it might use yet higher-order conjunctive nodes devoted to encoding the entire proposition ([('man' & *biter*) & ('dog' & *bitee*)]). In their simplest form, these models are analogous to models of object recognition that proceed by activating neurons or groups of neurons tuned to successively higher-order combinations of input features (Reisenhuber & Poggio, 1999).

Conjunctive representations can, in principle, exist at various levels of abstraction, and across representational content. These could, for example, be predicate-specific ("man as biter") (cf. O'Reilly & Busby, 2002), or represent more general semantic role/filler combinations ("man as agent"). (Doumas et al., 2008; Hummel & Holyoak, 1997; Hinton, 1986). It is worth noting, however, that any model of complex meanings represents conjunctions. Whenever a variable is assigned a value, it represents a conjunction of features. At issue is the strategy used to encode this combination. Models described as "conjunctive" are typically those with physical units devoted to conjunctions, separate from the physical units encoding the conjoined.

Researchers have established that simple conjunctive coding models that devote one, or a few nodes to representing each conjunction are inadequate, however (Marcus, 2001; Pinker, 1997, Hummel & Holyoak, 2000; 2003). To their credit, most modelers that deploy conjunctive coding also recognize it as insufficient for composition, using it instead for a narrower function, such as long term memory storage of bindings (Hummel & Holyoak, 1997; Doumas et al., 2008). Its shortcomings are nonetheless instructive, and worth elaborating. For, even if simple conjunctive coding models are straw men, they are straw men that can scare off crows hovering above better models.

First, simple conjunctive coding models have difficulty encoding new role-filler combinations (Marcus, 2001). Humans do not. To encode a new combination, the model needs a stock of unassigned neurons, and a procedure for using those neurons to encode new combinations on the fly. It is unclear how a conjunctive model could identify neurons that are representationally unspoken for, and assign them to encode bindings as needed. This is the problem of generalization for conjunctive models. However, to be fair to conjunctive models, their neuron allocation problem is endemic to all of the cognitive neuroscience of memory, including episodic memory. Computational solutions have recently been proposed that may allow neurons in the hippocampus to be flexibly allocated to encode new conjunctions (Valliant, 2012).

But, even if this particular problem is ultimately tractable, the situation is still dire for simple conjunctive coding models. For, perhaps even more fundamentally, it's well established that simple conjunctive models would require more neural resources than the brain has available (Pinker, 1997; Marcus, 2001). In the simplest case, a neuron would be needed for every unique sentence a person ever understood or produced, and every unique thought one ever entertained. This is the problem of scale, and it recurs throughout our consideration of alternative models of binding, given the math of combinatorics. This is a fundamentally different type of problem than problems with (e.g.,) generalization. It is a problem in physical implementation, not cognitive adequacy. We hold models to both standards throughout this review.

Researchers have attempted to mitigate both the issue of generalization and the issue of scale by encoding conjunctions using coarse, distributed representations. Rather than encoding each conjunction using one or a few nodes, models can instead use and reuse the same nodes, encoding different conjunctions as unique activity patterns over the same set. O'Reilly and Busby (2002) propose such a model for encoding spatial relations (e.g., "the triangle is above the circle"), and show that this model can encode a large number of conjunctions using relatively few nodes. The model has one large bank of nodes that encodes conjunctions for all the various spatial relations. There is no internal differentiation of the representation of the different variables, as such, separate from the representation of their values. Using this scheme, the authors are able to encode 62,400 conjunctions using only 200 nodes. This is a meaningful step toward dealing with the combinatorial explosion that accompanies conjunctive coding. Moreover, distributed representations allow the network to represent new conjunctions on the fly, encoding novel combinations as a function of their similarity to old combinations. This distributed model is thus an improvement over simple, localist conjunctive models, with cardinal neurons. It mitigates the issue of scalability, and provides the network some ability to generalize to new role/filler combinations. Nonetheless, it has major limitations.

Putting aside global concerns about connectionist models (biological implausibility due to bidirectional activation flow, heavily supervised learning, requiring an impossible amount of training experience (see e.g., Fodor & Pylyshyn, 1988;), there are a number of problems with conjunctive models that are not remedied by the use of coarse, distributed representations. The first is a minor, but overlooked, problem that is specific to the types of representations O'Reilly and Busby use. The second two hamper all conjunctive coding models, whatever their representational scheme.

O'Reilly and Busby's scheme works because the same nodes are called upon repeatedly, each time encoding one small piece of the information necessary to represent the entire conjunction. But there are biophysical drawbacks to implementing this solution in the brain. This is because, contrary to common assumption, theoretical work suggests that coarse, distributed representations would impose steep metabolic costs on the brain's operation. It has been estimated that up to 80% of the brain's energy is consumed by neural signaling, with the remaining percentage going to processes required for building and maintaining cells (Sibson et al., 1998; Rothman et al, 1999; Laughlin, 2001). Under such a ratio of signaling to fixed energy costs, Laughlin (2001) estimates that the most metabolically efficient codes are distributed, but sparse codes; codes that involve numerous cells per bit of information encoded, but in which each cell fires only infrequently. Very coarse, distributed representations, like the one's used by O'Reilly and Busby may thus be a drain on the brain's metabolic resources, relative to other signaling strategies (Lennie, 2003; Laughlin, 2001), given that they require a very large number of active neurons to encode some piece of information. Of course, since O'Reilly & Busby's model is specified at the abstract level of nodes rather than neurons, we cannot properly evaluate whether it's plausible to implement. It's unclear how exactly the activity of a node in a neural network is mapped on to the firing rate of neurons at the neural level. But we should nonetheless be wary of coarse, distributed models that require a large number of active nodes to encode some content that could, in principle, be encoded by less neural signaling. Energy consumption considerations suggest that distributed representations are useful, but only up to a point. Though using and reusing the same nodes may reduce the number of nodes needed, it may not actually minimize energy consumption, all things considered. It's not obvious that O'Reilly and Busby could simply reduce the number of nodes per bit of information, and still maintain the network's ability to generalize to novel conjunctions. This, of course, is not to say the brain could not, in any circumstance, implement such a model. But it appears to be a poor solution from the perspective of minimizing energy consumption, relative to a fixed task goal. In assessing the feasibility of a strategy, it is thus important to not only consider how many neurons are required to implement it, but also how much activity is required of those neurons.

Even if a more suitably energy efficient coding solution were found, however, there are deeper problems with conjunctive models. These are problems with their adequacy as cognitive models, rather than the plausibility of implementing them. First, there is no straightforward way for the network to signal which bindings are part of the same proposition (Pinker, 1997; Jackendoff, 2002). This is especially clear when we consider cases in which the same role occurs twice. Take as an example the sentence "John loves Mary, but Mary loves Jesus". Here, Mary is both a lover and a lovee, but in different propositions. If the network simply encoded first-order conjunctions, it knows John is a lover, Jesus is a lovee, and Mary is both lover and a lovee. But it's terribly unclear who loves whom. Does John love Jesus? Does Mary love herself? These conjunctions need to be brought together within propositions in order for the mind to know who loves whom. The O'Reilly & Busby model does not take the extra step to encode entire propositions, and it's unclear if it could manage it in a computationally tractable way.

Second, although the network can generalize, there is no guarantee that it will generalize systematically (Fodor & Pylyshyn, 1988). That is, it's not a given that any network that can represent "Faust defeated Summers" can thereby also represent "Summers defeated Faust". Conjunctive models therefore seem to miss an essential fact about the mind. If the model happens to have encountered something similar to a square, say a rectangle, in the *left-of* role, then it may generalize appropriately. But this seems to lean too heavily on the model's idiosyncratic experience with particular classes of things in particular roles. We will see ways that rule-like binding procedures can be implemented in a neural network below.

Finally, and perhaps most importantly, the patterns representing, for example, 'dog' in 'dog bites man' appear to have no relationship to the pattern representing 'dog' in 'man bites dog'. There is a pattern of activity for "dog as biter" and a different pattern of activity for "dog as bitee", but no systematic relationship between the two. It's thus not clear how the identity of the representation 'dog' is maintained, and hence can contribute the same meaning to both complex expressions. Thus, compositionality is lost. Hummel et al. (2000) make essentially the same point, calling for "role-filler independence" in the encoding of bindings. The principle of "role-filler independence" simply holds that bindings must be created in such a way as to maintain the identity of both the roles and fillers (variables and values). This promotes compositionality, and also, they maintain, allows the network to learn things about the roles and fillers, as such, across different combinations involving them.

Temporal Synchrony as a Binding Signal

To some, these concerns weigh decisively against any type of conjunctive model (Hummel et al., 2000; Doumas et al., 2008), and more generally, against any model that brings variable and value together in the same neuron or neural population (Hummel et al., 2000). These writers and others favor a fundamentally different physical signal for binding: temporal synchrony between disparate neural populations (von der Malsburg, 1994; 1999 Ajjanagadi & Shastri, 1991; Doumas et al, 2008; Holyoak & Hummel, 2000; Love, 1999; Singer & Gray, 1995). Synchrony-based models signal the binding of a value to a variable through the transient synchronization of the units encoding variables and those encoding the values. Thus, to encode "man bites dog", the "man" units would become transiently phase-locked with the "biter" or "agent" units. Meanwhile, the units encoding "dog" are entrained with the "bitee" or "patient" unit. The two couplings then fire in different phases ("man" and "biter" in one phase, "dog" and "bittee" in another phase) distinguishing the two bindings for any downstream neurons that are interested. Note, however, that synchrony is only meant as a signal, or tag, for what is bound to what. It plays no role in the computations determining the binding. The computations driving which representations get synchronized are performed by an independent group of neurons.

At first glance, synchrony has a number of desirable properties as a representational strategy. It appears to avoid the combinatorial explosion of neurons required to encode conjunctions. A mind that binds through synchrony needs only as many units as is necessary to represent the set of fillers and the set of roles. These can then be flexibly and dynamically combined at no additional imposition on cortical space, dealing neatly with the problem of scale that plagued conjunctive coding. Second, both fillers and roles contribute the same content whenever they are used. The representation of Joe Biden is the same in "Joe Biden is the vice president of the United States", "Joe Biden spent Tuesday morning washing his Chevy Camero", and "Joe Biden bit the dog". Likewise for the content of the roles. In this, temporal synchrony appears more promising as a means of achieving compositionality than the conjunctive strategies discussed above.

But, on closer inspection, temporal synchrony may not actually reduce the number of neurons required at all. The synchrony code is useful only if it can be read by other neurons. Von der Malsburg (1999) argues that synchrony-based codes can be read by "coincidence detector" neurons that receive input from both variable and value neurons. Shadlen and Movshon (1999) point out that such coincidence detectors amount to little more than the conjunctive neurons synchrony was supposed to abolish, however; they are neurons that fire if and only if both constituent neurons are active at the same time, and in the same phase. The brain therefore needs just as many coincidence detectors as is necessary to represent all the possible coincidences that need to be detected; that is, as many as necessary to encode the conjunctions of features.

If additional neurons first compute which neurons fire in synchrony, and further neurons read out which neurons are currently synchronized, why not eliminate the middle neuron? Synchrony is playing no obvious functional role. The coincidence detector neurons could simply be connected to the neurons computing the bindings, with no loss in information. The challenge, as before, is to set these up in a way that allows the network to systematically encode new combinations, and do so within the constraints of the brain's physically available resources.

Classical Symbol Structures in Neural Networks

Given the ease with which classical cognitive architectures explain compositionality, perhaps it should be no surprise that the cleanest architectural solutions to this problem are inspired by the digital computer. In this spirit, Marcus (2001) proposes that the human brain has registers, analogous in function to the registers of a digital computer, that represent the values of particular variables. Different registers could represent different variables (e.g., agent/patient), thus tying the identity of the variable to the identity of the register. Symbols representing these values could be stored in these registers to encode a variable/value binding. One important aspect of this model is that the same symbol, say a string of bits, or even the address of a symbol in long-term memory, can be stored in different registers. Thus, the symbol contributes the same content across tokens, and the meaning of a complex expression is a function of the meaning of its parts. Both the fillers and their roles maintain their independence, per Hummel et al.'s (2000) requirement, preserving one advantage we saw for temporal synchrony models over simple and distributed conjunctive coding models. However, there is no additional binding signal needed over and above the assignment of a symbol to a register (value to a variable).

Although this model is specified at the representational level, it's certainly logically possible to implement this and related architectures in neural networks. But can it be implemented without falling victim to the difficulties of the conjunctive models reviewed above? To make a fair comparison to the other models, we should consider the plausible ways in which to implement this in a neuron-like substrate.

A broadly similar, if limited, solution can actually be found in Hinton's (1986) toy neural network model for encoding family relations. This model dedicates separate banks of nodes to representing different variables: one bank encodes the first argument of the relation, a second encodes the relation, and a third encodes the second argument of the relation. Yet another layer encodes the entire proposition. The values of each of these variables are encoded by distributed patterns of activity across these banks of nodes. Despite their superficial differences, Hinton's model is similar to Marcus's suggestion in that it uses spatially individuated sets of nodes to encode different variables (akin to registers), and unique distributed codes within these registers to represent those variables' values[2]. The same code can be used to represent a symbol's identity across node-banks. Hinton's model is limited in the novelty of the bindings it can encode because it is trained through back-propagation, and hence dependent on the similarity of a new combination to previously experienced combinations. This problem can be remedied by simply abandoning the unnecessary commitment to back-propagation, however.

For example, Smolensky' s (1990) tensor product model provides a very general procedure for combining two pieces of information, coded as vectors, into a single complex representation, coded as a matrix. Here, variables and values are represented as patterns of activity over different set of nodes. These patterns are thus distinct vectors. If one vector codes for a variable (e.g., agent) and another codes for a value (e.g., man), their binding can be represented as the element-by-element multiplication of the two vectors. The product of this multiplication is called a "tensor product", and this output is represented on a set of nodes distinct from the nodes encoding the variables and values. The tensor product of an *m* dimensional vector and an *n* dimensional vector is an *m*X*n* dimensional matrix, which

[2] Strictly, Hinton's would not qualify as a register model, *per se,* because it doesn't have memory for the patterns of activity instantiated across the role units. However, this seems like an inessential feature of the model, as recurrent connections could be implemented to achieve this.

encodes the binding. This circuitry implements a rule-based procedure that can take any variable and any value and bind them, even if that combination has never been encountered, thus dealing with the problem of generalization that accompanies models trained with back-propagation. This binding is not limited to two vectors, however. First-order bindings (e.g., "man bit" and "bit the dog") can thus, in principle, be combined into a proposition ("the man bit the dog"). These propositions can then be combined to form increasingly higher-order propositions. There is therefore no principled problem in encoding "John loves Mary, but Mary loves Jesus," as there is in other conjunctive coding models.[3]

Hummel et al. (2000) maintain (and others have echoed (Haysworth, 2012; Kriete et al., 2013)) that the tensor product violates the principle of role-filler independence. They argue that once two vectors are multiplied to encode some complex meaning, there's little sense in which 'dog' contributes the same meaning to "man bites dog" as "dog bites man". The 'dog' pattern is now inextricably bound with different role-patterns, and represented on separate physical units than the basic symbol 'dog', whose meaning is supposed to be static. Whether this is a legitimate concern, however, depends on whether there are procedures for faithfully recovering the constituent vectors ('man', 'dog', 'bites') from the combined vectors ('the man bit the dog'). Smolensky claims this unbinding is possible if the patterns used to represent either the contents of the set of roles or the set of fillers are linearly independent (i.e., uncorrelated). Linear independence is a strong requirement, but

[3] There may, however, be a problem with scale (Plate, 2004; Marcus, 2001). As representational structures get larger, ("I believe that John loves the book that Alice wrote") the size of the network grows exponentially. Smolensky argues that this may mirror human difficulty with complex structures, though whether there is indeed such a model-mind correspondence is unclear. Plate (1994) introduces a circular convolution procedure that can be used to keep the size of the network fixed across increasingly large bindings.

it is not at all implausible if the number of roles is small. In the simplest case, one or a small number of nodes could be uniquely devoted to each role, ensuring that they are independent. For example the vector [1 0] could represent the agent and [0 1] the patient. This would allow the constituent representations to be recovered from the complex representations in a law-like way, mitigating Hummel et al.'s criticism.

Although dedicating separate neural populations to explicitly represent different semantic variables is somewhat antithetical to the connectionist spirit of Smolensky's proposal, it emerges as an appealing solution. Here, it guarantees that the identity of the roles and fillers are independently recoverable, and that the symbols that enter into bindings maintain some semblance of their identity. Unbinding a tensor product would also be possible if the vectors that encode the fillers, and not the roles, are linearly independent. Given that there are many more possible values that can fill a given variable than there are variables, this would be more difficult to implement and maintain linear independence. However, there may be good reasons for the brain to adopt, or at least approximate, this strategy. We begin to address this empirically in Experiment 3. Whether the tensor product is compositional thus depends upon the particular representational scheme the brain uses for encoding the variables and variables.

There is, however, one final solution that been proposed, again based on the classical computer, that can clearly preserve compositionality. In order to ensure that a symbol has the same identity across bindings, models can add an additional layer of representation, mimicking the pointers of classical computers. Kriete et al. (2013) adopt this approach, implementing variable/value binding in a neural network using a classical pointer system. This neural network binds variables and values by temporarily storing the address of a symbol in a variable-specific memory register. The architecture is thus functionally identical to the pointer systems used by digital computers, which represent the address of a symbol in the system's long-term memory. This "indirection", as it's known in computer science, preserves the identity of the symbol across roles, mitigating Hummel et al.'s concerns about role-filler independence. In this, it effectively implements Marcus's (2001) suggestion in a neural network. These registers could, in principle, be hierarchically organized in sets in order to encode multiple bindings, entire propositions, and propositions within propositions. Kriete's model does not take these additional steps, but there seems to be no problem, in principle. We address the concern that such a model would require too many neurons to implement in the General Discussion.

What is most important, for present purposes, is the general representational architecture of Kriete et al.'s model (and perhaps the modified tensor product described above). To date, this is the best neural network model for encoding structured meaning. Perhaps not coincidentally, it is also the most classical in its flavor. It explicitly represents variables as such, and allows their values to be flexibly updated. It can bind new values to variables. And, although Kriete et al. don't emphasize this aspect, other cortical systems can easily reference these variables in their operation, because the variable registers are spatially individuated. Operations that need to know who the agent is need only be able to decode the pattern of activity in that register. Is this how the brain actually does it?

Where in the brain might structured meanings be represented?

We saw above that many theoretical models encode structured meanings by explicitly representing abstract variables that can take different values ("who did it?, "to whom was it done?") (Marcus, 2001; Hinton, 1986; Smolensky, 1990; Kriete et al., 2013). Moreover, the best of these models use different neural populations to separately encode the values of each of these variables. Of course, how the brain actually accomplishes this is not just a theoretical issue. We should, in principle, be able to identify these neural representations, and test whether the brain represents these variables, and how. Fodor & Pylyshyn (1988, pg. 13) write: "…the symbol structures in a Classical model are assumed to correspond to real physical structures in the brain and the combinatorial structure of a representation is supposed to have a counterpart in the structural relations among physical properties of the brain". Here, we ask, what and where are these physical properties?

Although there has been little direct work on how the brain represents complex, structured meanings, the study of the neural basis of sentence comprehension, more generally, has a rich history. This literature provides clues for which regions we might expect to represent who did what to whom. Compositionality requires both combinatorial mechanisms and basic symbolic content, and this literature has targeted the syntactic and semantic combinatorial mechanisms. Neural systems that compute a representation of the syntactic structure of sentence, and those that then map that representation to a structured semantic representation are both relevant here.[4]

A reliable network of cortical regions surrounding the left sylvian fissure are involved in building and interpreting structured linguistic representations, over and above

[4] We assume that both syntactic and semantic representations can be complex and structured. That is, they can both have simpler representations as parts, and, in both cases, the way in which the parts are assembled determines the identity of the complex representation. The theoretical issues outlined in the introduction therefore apply to both. We assume that semantic composition proceeds from a representation of the sentence's syntax. Throughout, we are not always able to say where the representations we identify in fall with respect to this division. For consistency and clarity, we consider them semantic, or more specifically "structure-dependent semantic representations": meanings that depend upon a structured syntactic representation of the sentence.

word-level linguistic representations. This sentence-processing network includes portions of the inferior frontal cortex (Hagoort et al., 2004; Fedorenko et al., 2011; Pallier et al., 2011; Meltzer et al., 2009; Dapretto & Bookheimer, 1999; Bornkessel et al., 2007; Grodzinsky & Friederici, 2006; Kuperberg et al., 2000; Hagoort & Indefrey, 2014), inferior parietal lobe (Fedorenko et al., 2011; Pallier et al., 2011; Meltzer et al., 2009; Humphries et al., 2006) much of the superior temporal sulcus and gyrus (Bornkessel et al., 2007; Dronkers et al., 2004; Fedorenko et al., 2011; Pallier et al., 2011; Meltzer et al., 2009; Devauchelle et al., 2009; Vandenberghe et al., 2002; Humphries et al., 2006; Friederici et al., 2003; Bornkessel et al., 2007), and the anterior temporal lobe (Meltzer et al., 2009; Bemis & Pylkkanen, 2011; Baron & Osherson, 2011; Devauchelle et al., 2009; Rogalsky & Hickok, 2009).

Many neuroimaging studies have localized this network by contrasting the magnitude of activation when reading or listening to meaningful sentences to the magnitude of activation when processing unstructured word lists (Mazoyer et al., 1993; Vandeberghe et al., 2002; Humphries et al., 2006; Fedorenko et al., 2011). It's not surprising that this particular contrast would activate a large network, given that these tasks differ in the presence of local and global syntactic structure, the assignment of thematic roles, the construction of a semantic representation of an event, as well as any inferences about that event. However, neuroimaging paradigms that have more tightly controlled specific syntactic or semantic factors in order to precisely localize specific functions have also met with varied results.

For example, rather than using a cognitive subtraction approach, Friederici et al. (2003) import an anomaly-detection paradigm from the electroencephalogram literature

(Kutas & Hillyard, 1980; Hahne & Friederici, 1999) hoping to differentially localize structured semantic and syntactic processes. In this paradigm, subjects listen to sentences fraught with either semantic or syntactic violations ("the thunderstorm was ironed"/"the blouse was on ironed"). If a brain region responds more to semantic than syntactic anomalies (or vice versa), it is inferred to encode semantic information (or vice versa, syntactic information). Using this paradigm, Friederici et al. (2003) found regions of left and right mid-superior temporal gyrus, as well as the bilateral insula to respond more to semantic than syntactic anomalies. The left-posterior STG responded robustly to both anomaly types relative to correct sentences, and the anterior STG responded more to syntactic than semantic violations. This approach may therefore seem promising in winnowing down the large number of perisylvian regions listed above that have been identified by coarser contrasts. However, using very similar paradigms, Hagoort et al., (2004) and Kuperberg et al. (2003) also report increased activation to semantic anomalies in the left inferior frontal cortex. Further, Kuperberg et al., (2003) expand the list of superior temporal regions that respond more to semantic than syntactic anomalies to include the entire length of the superior temporal sulcus, from posterior superior temporal sulcus to the temporal pole. Thus, the majority of the regions constituting the perisylvian sentence-processing network appear to be modulated by the detection of phrase or sentence-level semantic anomalies. These results thus converge with cognitive subtraction contrasts regarding which brain regions sub-serve structure-dependent semantic processing. But, when taken together, they do not reliably fractionate the perisylvian network into specific representational sub-components.

Rogalsky & Hickok (2009) worry that these anomaly-detection paradigms could recruit specialized error-detection mechanisms, rather than simply reflecting normal semantic and syntactic processing. To deal with this concern, they adopt a modified anomaly detection paradigm. The authors had subjects selectively monitor visually presented sentences either for semantic violations ("that nest in the robin was laying eggs") or syntactic violations ("those athlete in the gym was lifting some weights"). However, instead of measuring and comparing activity to different types of anomalous sentences, the authors only analyzed data from normal, non-anomalous, sentences. Any differences in activity therefore reflect differences in top-down attention when attending to alternate classes of representational content (semantic vs. syntactic). Using this paradigm, they found a large area of the left anterior temporal lobe (ATL) that responded to the sentences, and a very small sub-area of left ATL that was modulated only by attention to semantic anomalies. The authors suggest that this region of left ATL is selective for compositional semantics, given the intermingling of structured syntactic representation and attention to structured meaning. However, one could voice a similar concern to that Hickok & Rogalsky themselves voiced about anomaly-detection paradigms: it's not obvious that the higher-order representations that encode whether one is monitoring for semantic or syntactic anomalies are directly related to the actual semantic and syntactic representations themselves. Monitoring for semantic errors and constructing the meaning of a sentence are simply different processes.

Nonetheless, the conclusion that anterior left ATL is selectively involved in building compositional semantic representations finds support in other studies that do not suffer from this shortcoming. For example, Pallier et al (2011) developed a well-controlled paradigm in which they could vary the number of syntactic constituents in a sentence, and these constituents could be syntactically well formed and meaningful, or syntactically well-formed but meaningless jabberwocky. They found a large region of left ATL that monotonically tracked the number of constituents in a sentence, but only when those constituents were meaningful. The ATL was not the only such region, however. The authors also found regions of left mid-STS and the left inferior parietal lobe that show the same interaction. Areas of the posterior superior temporal sulcus and inferior frontal cortex tracked the number of constituents in the sentences regardless of whether or not they were meaningful. We thus again find clear evidence that regions of the perisylvian language network are involved in building semantic representations, given syntactically structured input. In this case, the ATL, mid-anterior STS, and inferior parietal lobe appear to integrate semantic information over the sentence. However, we also find that numerous sub-regions therein appear to play functionally similar roles.

This general pattern of broad coherence in the perisylvian network, with little reliable internal differentiation is not merely attributable to functional neuroimaging. Dronkers et al., (2004) report convergent results in a large study of patients with various types of brain damage. These patients were asked to match sentences to pictures depicting their meaning. They did so for many different sentence-types, ranging from simple declaratives (e.g, "the girl is sitting"), to reversible active sentences ("the girl is pushing the boy"), to a number of more syntactically complex construction types such as, for example, those involving object clefting ("It is the boy that the girl chases"). The authors then examined where brain damage predicts impairment on each sentence-type. The most notable finding was that a region of the posterior left middle temporal gyrus strongly predicted comprehension deficits across all sentences types. The authors suggest that this region is involved in word-level comprehension. Beyond this MTG region, the authors report that damage to the anterior superior temporal gyrus and superior temporal sulcus, posterior regions of the superior temporal sulcus and inferior parietal cortex, and regions of the left inferior frontal cortex all predict broad impairments in sentence comprehension. We review more specific findings of this study in the General Discussion. For now, we merely note that this lesion study identifies the same perisylvian sentence processing network as the neuroimaging work, and again finds little differentiation between the sub-regions of this network. MTG is the one region that shows a unique pattern of deficits, and its deficits appear to be at the word rather than phrase or sentence level.

As a result, the literature contains credible proposals that at least some semantic representations with compositional structure are built in the left inferior frontal cortex (Hagoort et al., 2004; Hagoort, 2007; Hagoort & Indefrey, 2014)), the left anterior temporal lobe (Vandeberghe et al., 2002; Hickok & Poeppel, 2007; Westerlund et al., 2015), and the left parietal lobe (Binder et al., 2009; Humphries et al., 2006). From this literature, one might conclude that many regions around the sylvian fissure play the same role in building structured linguistic representations. However, a tradition of neuropsychological work has shown repeatedly that brain damage can cause remarkably specific functional deficits in language comprehension (Caramazza & Zurif, 1976; Schwartz, Saffran, & Marin, 1980; Linebarger et al., 1983; Caramazza & Miceli, 1991). Caramazza & Micelli (1991) report a patient with parietal lobe damage who has selective inability to assign thematic roles, but intact morphosyntactic processing involving closed-class elements. Linebarger et al. (1983) find that the ability to assign thematic roles can be impaired, while general judgments of the grammaticality of a sentence remain intact. Thus, there is a minor tension between the functional modularity demonstrated in neuropsychology, and the relative lack of functional specialization reported (in aggregate) in the perisylvian network, especially as assessed in functional neuroimaging. It thus seems more likely that the difficulty in reliably fractionating the perisylvian network lies in difficulty establishing reliable mappings between structure and function across subjects. This may result from deficiencies in the conventional ways of aggregating data across subjects (See Fedorenko et al., (2010; 2011) for alternative strategies for identifying functional ROIs in individual subjects that may prove fruitful in this regard), or perhaps from functional contrasts that have been too coarse.

The present work

From this literature, we expect regions surrounding the left sylvian fissure to construct meaning from syntactically structured input. But how? Here, instead of asking where the magnitude of neural activity is modulated by the presence of syntactic and/or semantic combinatorial operations, we ask simply: what brain regions carry information about a sentence's meaning? And then how is this meaning encoded? In this, we seek the "physical counterpart" to the combinatorial structure that Fodor & Pylyshyn (1988) suggest exists, if the mind has a classical architecture.

To do so, we use fMRI, in conjunction with Multi-Voxel Pattern Analysis (MVPA). The development of MVPA has dramatically improved researchers' ability to identify the representational content of brain states using fMRI (Haxby, 2001; Norman et al., 2005; Mitchell et al., 2008). Rather than asking whether the mean amplitude of a brain region's response is modulated by an experimental contrast, researchers now routinely pool information across voxels to directly assess that region's informational content. The brain, as the seat of the mind, is an information-processing machine. Information isn't just contained in the peaks and valleys of neural activity, but in the undulations, however small, of the entire landscape. Testing for differences in informational content, rather than magnitude of activation, allows researchers to identify much subtler cognitive and representational distinctions; distinctions that aren't reflected in any aggregate difference in response amplitude (Kamitani & Tong, 2005; Eger et al., 2008; Kriegeskorte et al., 2007). If pattern classification algorithms can reliably identify the content of a representation given a pattern of BOLD signal, then other brain regions can likely decode these representations as well.

MVPA is thus particularly useful in the present context. There are clear differences in the content of the propositions conveyed by "the truck hit the ball" and the "the ball hit the truck", respectively. Likewise for "Faust defeated Summers" and "Summers defeated Faust" etc. However, this is not a difference that we would expect to be manifest in gross differences in the response amplitude of an entire region. It isn't as though the construction of one of these propositions engages a cognitive process that the other doesn't. The same combinatorial mechanisms and same symbolic content should be engaged in both cases, only with a different functional mapping between the roles and fillers.

In the remainder of the thesis, we describe three fMRI experiments aimed at understanding how the brain (in the left perisylvian cortex or elsewhere) flexibly encodes the meanings of sentences involving an agent ("Who did it?"), an event-type ("What was done?"), and a patient ("To whom was it done?").

Experiment 1: Where is who did what to whom?

In order to study how the brain represents structured meanings, we need to know where in the brain these representations are located. Here, we look for regions that exhibit different patterns of activity for different propositions assembled using the same concepts (“the truck hit the ball”/”the ball hit the truck”). That is, we look for brain regions from which we can decode who did what to whom?

In experiment 1, subjects undergoing fMRI read reversible sentences describing simple events. Reversible sentences have a long and productive history in the neuropsychology of language. They have been valuable tools in probing the various patterns of functional deficits that can underlie so-called “agrammatic” aphasia (Caramazza & Zurif, 1976; Schwartz, Saffran, & Marin, 1980; Ansell & Flowers, 1982; Caramazza & Miceli, 1991; Berndt et al., 1996). In the present work, we put reversible sentences to a slightly different use. Each reversible sentence expressed a meaning, or proposition, that could be conveyed in either the active or passive voice (e.g., “the ball hit the truck”/“the truck was hit by the ball”). Each such sentence could be reversed to yield a mirror-image proposition (e.g., “the truck hit the ball”/“the ball was hit by the truck”), which was also included in the stimulus set. We call these “mirror-image proposition pairs”. The use of mirror-image propositions ensures that basic lexico-semantic content, syntactic tree structure, and summed word frequency are matched within pairs. Members of these pairs contain the same words and have the same syntactic structure, but the words are differentially assigned to the agent and patient roles to form different propositions.

A region encoding these meanings should have the following two properties. First, patterns of activity in such a region should differentially encode members of mirror-image propositions pairs. For example, the propositions conveyed by “the truck hit the ball” and "the ball hit the truck" should elicit distinct patterns of activity. Second, the instantiation of such patterns should predict downstream neural responses that depend on understanding who did what to whom. For example, patterns related to structure-dependent meaning should predict differential affective responses to "the grandfather kicked the baby" and "the baby kicked the grandfather". Experiment 1 employed two key analyses, corresponding to these two functional properties. First, we applied MVPA and a whole-brain searchlight procedure (Kriegestkorte et al., 2005) to identify sets of contiguous voxels that distinguish between members of mirror-image proposition pairs. Second, we developed a pattern-based effective connectivity analysis (PBEC) to determine whether patterns related to affectively salient sentences (e.g. "the grandfather kicked the baby") mediate the relationship between the sentence presented and affective responses elsewhere in the brain. Jointly, these analyses establish candidate regions for encoding structure-dependent meaning that can be further probed in Experiments 2 and 3.

Wholebrain Searchlight Analysis. First, using a linear classifier, we searched for regions whose patterns of activity distinguished between members of mirror-image proposition-pairs: for example, between the proposition conveyed by "the truck hit the ball" (as well as "the ball was hit by the truck") and the proposition conveyed by "the ball hit the truck" (as well as "the truck was hit by the ball"). Active and passive forms of each proposition were treated as identical in all analyses, targeting underlying semantic representations, controlling for visual features of the stimuli and surface syntax. All propositions were presented separately, and multiple times, to better estimate the pattern of activity evoked by each proposition. For Experiment 1, classifiers were thus tested on their ability to discriminate between new tokens of the mirror-image propositions on which they were trained.

For this initial searchlight analysis, we used four mirror-image propositions pairs, two involving animate entities and two involving inanimate entities. For each subject (N=16), we averaged classification accuracies across these four pair-wise classification problems to yield a map of the mean classification accuracy by region. Group-level analysis identified a region of left-mid superior temporal cortex (lmSTC, k=123, Talairach center: -59, -25, 6) that reliably distinguished between mirror-image propositions (*p*<0.0001, corrected. See left temporal region in Figure 1). We find fairly consistent levels of classification accuracy in both ROIs, suggesting these regions are not driven by idiosyncrasies of particular pairs, or only by the animate or inanimate proposition-pairs. (See Table 1). A repeated-measures ANOVA revealed no significant differences in classification accuracy across the six pairs for lmSTC (*F*(5, 75)=0.4, *p*=0.84), and all are significantly above chance, or trending. The pattern of results is consistent with lmSTC encoding domain-general information about who did what to whom for this class of events.

lmSTC	**mirror-image pair**	**Mean Accuracy**	***t*, one-tailed *p***
1	STRUCK (mother: boy)	.566	*t*=1.8, p=0.046
2	KICKED (grandfather: baby)	.602	*t*=3.83, p=0.0008
3	TOUCHED (grandmother: girl)	.563	*t*=1.45, p=0.0843
4	PULLED (father: child)	.565	*t*=1.89, p=0.039
5	HIT (truck: ball)	.556	*t*=2.5, p=0.012
6	SMACKED (door: branch)	.597	*t*=4.3, p=0.0003

Table 1. Post-hoc analysis delineating classification by mirror-image proposition pair. Note, the first two propositions were not used in localizing the ROI, only in connectivity analyses.

A second significant cluster was discovered along the right posterior insula/extreme capsule region (*p*<0.001, corrected; -37, 9, 6)). However, this second region failed to meet additional, minimal functional criteria for encoding sentence meaning. Specifically, if the regions discovered in the searchlight analysis do represent structure-dependent meaning, then they should facilitate classification of non-mirrored propositions as well. For example, they should be able to distinguish "the truck hit the ball" from "the father pulled the child." Although these non-mirrored pairs are not well matched, and one would expect many other brain regions to be able to perform this classification (e.g., regions that encode the semantic/phonological content of the nouns and verbs), this analysis nevertheless serves as a sanity check on the ROIs localized using the searchlight analysis. Of the two ROIs able to discriminate within mirror-image proposition pairs, only the lmSTC ROI was able to reliably discriminate non-reversed pairings as well, t(15) = 4.06, *p*=0.005. The right extreme-capsule/insula ROI trended in this direction, but its results were not statistically significant, t(15) = 1.45, *p*<0.09. This failure to robustly classify non-mirror-image proposition pairs casts doubt on the possibility that the right posterior insula encodes complex, structured semantic representations.

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

To test this hypothesis, we used, within the same experiment, an independent set of mirror-image proposition pairs in which one proposition is more affectively salient than its counterpart, as in “the grandfather kicked the baby” and “the baby kicked the grandfather”. Differences in affective salience were verified with independent behavioral testing (See Appendix). We predicted that patterns of activity in lmSTC (as delineated by the independent searchlight analysis) would statistically mediate the relationship between the sentence presented and the affective neural response, consistent with a causal relationship (Baron & Kenny, 1986). To succeed, the mediating variable (here, the pattern of activity in lmSTC) must explain unique variance in the downstream response over and above variance in that response explained by the stimulus. This is because, to the extent that there is variability or “error” in this causal process, the more proximate mediating variable (here, the pattern of activity) should explain unique variance in the response, in virtue of being a channel through which the direct effect (here, the effect of the proposition presented on the amygdala’s activity level) is carried. This pattern-based effective connectivity (PBEC) analysis proceeded in three steps.

First, we confirmed that patterns of activity in the region of lmSTC identified by the searchlight analysis can discriminate between these new mirror-image propositions (t(15)= 3.2, p=0.005, mean accuracy = 58.3%), thus replicating the above findings with new stimuli. Second, we identified brain regions that respond more strongly to affectively salient propositions (e.g., “the grandfather kicked the baby” > “the baby kicked the grandfather”). This univariate contrast yielded effects in two brain regions, the left amygdala (28, 7, -18) and the right superior parietal lobe (40, 67, 44), (p<0.001, corrected). Given its well-known role in affective processing (Phelps & LeDoux, 2005), we interpreted this amygdala response as an affective signal, and focused on this region in our subsequent mediation analysis. Third, and most critically, we examined the relationship between patterns of activity in lmSTC and the magnitude of the amygdala's response. The first of the above analyses shows that "the grandfather kicked the baby" produces a different pattern in lmSTC than "the baby kicked the grandfather" (etc.). If these patterns actually reflect structure-dependent meaning, then these patterns should *mediate* the relationship between the sentence presented and the amygdala's response on a trial-by-trial basis.

To quantify the pattern of activity in lmSTC on each trial, we used the signed distance of each test pattern from the classifier's decision boundary (See Appendix for detailed description of how distance from the classification hyperplane is computed). Intuitively, one may think of this distance as the classifier's "confidence" in its decision regarding the test pattern. According to our hypothesis, trials in which the pattern is most confidently classified as, for example, "the grandfather kicked the baby" should be trials in which the amygdala's response is most robust.

As predicted, the pattern of activity instantiated in lmSTC predicted the amygdala's response (t(15) = 3.96, *p*=0.0013), over and above both the mean signal in lmSTC and the content of the stimulus. The pattern of activity in the lmSTC explains unique variance in the amygdala's response, consistent with a causal model whereby information flows from the sentence on the screen, to a pattern of activity in the lmSTC, to the amygdala (Sobel Test, z = 2.47, *p*<0.013. See Figure 1) (Baron & Kenny, 1986). The alternative model reversing the direction of causation between the lmSTC and amygdala was not significant (z=1.43, *p*=0.15), further supporting the proposed model. Alternative non-parametric tests for mediation yielded comparable results. (See Appendix). The above analyses were repeated with the right posterior insula/extreme capsule ROI as the mediator. In contrast to lmSTC, all assessments of whether the pattern of activity in the right insula/extreme capsule ROI mediated the relationship between the sentences presented and the amygdala were non-significant (*p*>0.15), further suggesting that the right insula/extreme capsule ROI does not represent structure-dependent meaning.

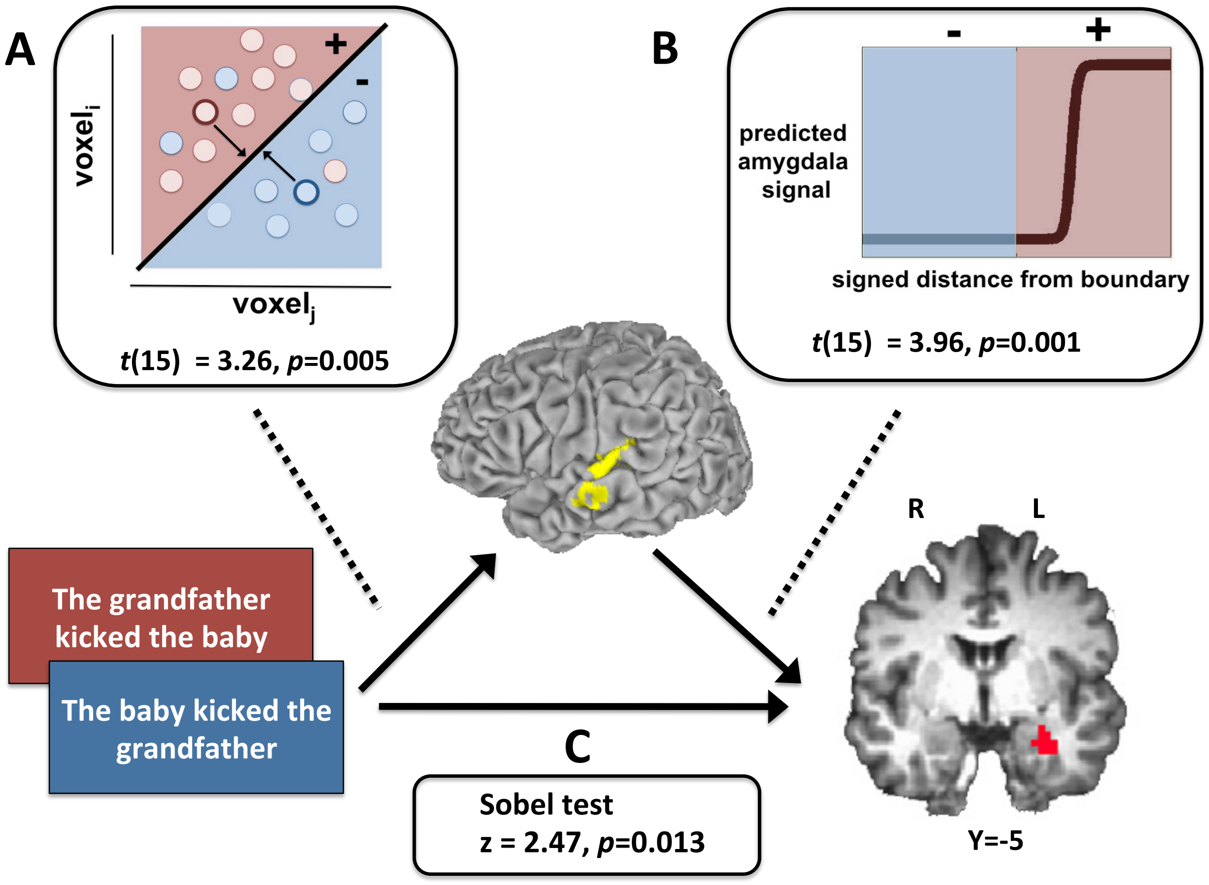
A
+
-
voxel$_i$
voxel$_j$
$t(15)$ = 3.26, p=0.005
B
-
+
predicted amygdala signal
signed distance from boundary
$t(15)$ = 3.96, p=0.001
The grandfather kicked the baby
The baby kicked the grandfather
C
Sobel test
z = 2.47, p=0.013
R
L
Y=-5

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

Thus, Experiment 1 shows that a region of lmSTC meets our two initial functional criteria for a region encoding structure-dependent sentence meaning. First, its patterns of activity differentiate between mirror-image propositions containing the same words and syntactic structure. Second, these patterns statistically mediate the relationship between the sentence presented and affective neural responses that depend on understanding “who did what to whom”. Previous research has implicated the left-mid superior temporal cortex in phrase and syntactic sentence-level semantic processing using both functional neuroimaging and lesion data (Friederici et al., 2003; Fedorenko et al., 2011; Meltzer et al., 2010; Pallier et al., 2011; Hagoort & Indefrey, 2014; Humphries et al., 2006; Vandeberghe et al., 2002; Dronkers et al., 2004; Wu, Waller, & Chatterjee, 2008). We save detailed discussion of the place of this finding within the literature until the general discussion, when we have a better understanding of this region’s functioning. Critically, experiment 1 leaves unspecified *how* this region encodes such information. Experiment 2 aims to further validate the results of Experiment 1 and to illuminate the mechanism by which this region encodes the meaning of these propositions.

Experiment 2: How does left-mid superior temporal cortex represent structured meanings?

We saw in the introduction that the most promising models of structure-dependent meaning represent the values of abstract semantic variables, allowing for the composition of complex meanings with constituent structure. In Experiment 2, we test the hypothesis that lmSTC flexibly represents these meanings (at least in part) by representing the values of the *agent* ("Who did it?") and the *patient* ("To whom was it done?"). To evaluate this possibility, we searched for sub-regions of lmSTC whose patterns of activity reflect the current value of these variables. We performed separate searches for each variable, looking for sub-regions of lmSTC that encode "who did it" and "to whom it was done" across verb-contexts. Thus, we aimed to identify regions that are specialized for representing the agent and patient variables *as such* (Marcus, 2001).

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

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

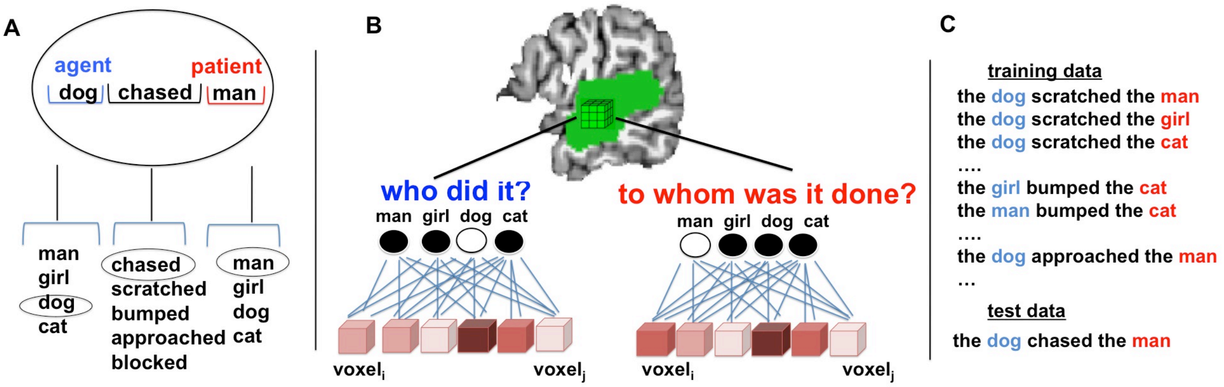

Figure 2. Experiment 2 design. **(A)** Subjects read sentences constructed from a menu of five verbs and four nouns, with one noun in the agent role and another in the patient role. **(B)** For each trial, separate pattern classifiers attempted to identify the agent and the patient based on activity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

These searchlight analyses revealed distinct sub-regions of lmSTC that reliably carry information about the identity of the agent and the patient (See Figure 3a). Within the anterior portion of lmSTC, a medial sub-region located on the upper bank of the superior temporal sulcus (STS) carried information about the identity of the agent (p<0.01, corrected; -46, -18, 1). A spatially distinct lateral sub-region, encompassing part of the upper bank of the STS, as well as the lateral superior temoral gyrus (STG) carried patient information_(*p*<0.0001, corrected, -57, -10, 2) across subjects. These anterior "agent" and "patient" clusters are adjacent, but non-overlapping in this analysis, and are significantly dissociable by their informational content ($F_{\mathrm{REGIONxROLE}}$(1, 24)= 12.99, *p*=0.0014. (See Figure 3.) This searchlight analysis also revealed a second "agent" cluster, posterior and superior to the clusters described above, located primarily within the posterior superior temporal sulcus (*p*<0.02, corrected, -57, -37, 7). Post-hoc analyses found the classification accuracies driving these results to be only modestly above chance levels of 25%, but statistically reliable across our set of 25 subjects (mean accuracies across subjects: anterior agent = 27.1%; posterior agent = 28.1%; patient = 26.6%).

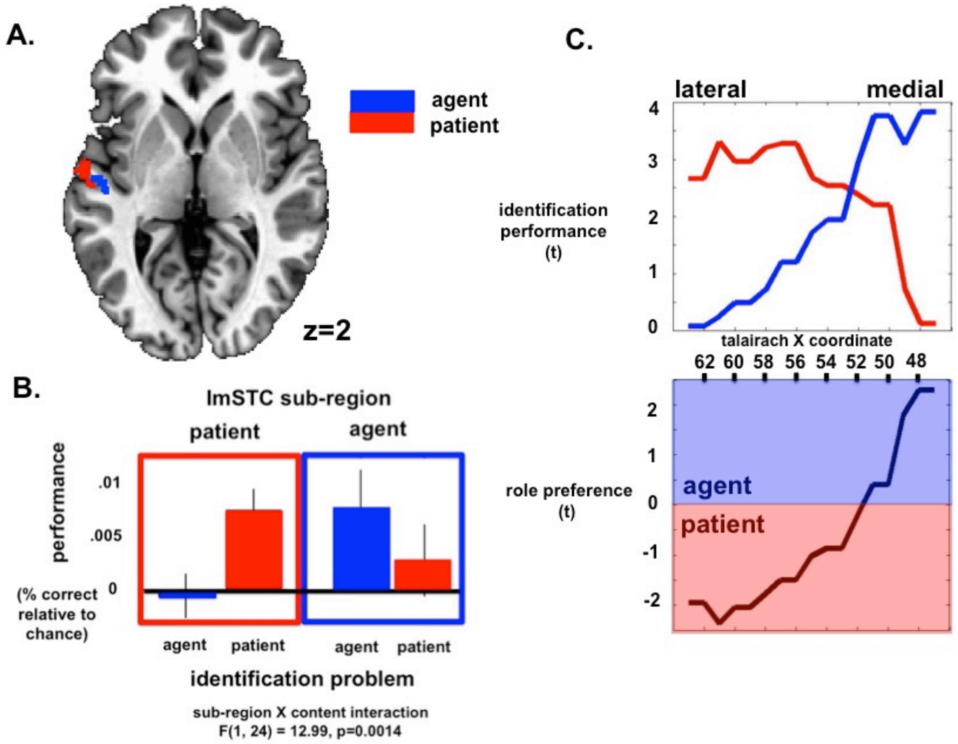

Figure 3. **(A)** Searchlight analyses identified adjacent, but non-overlapping sub-regions of anterior lmSTC that reliably encoded information about agent identity (medial, blue) and patient identity (lateral, red). **(B)**

Decoding accuracies of these regions, localized within each subject using independent data from other subjects, confirm that these adjacent regions differ significantly in the information they encode. **(C)** Across subjects, medial slices of anterior lmSTC preferentially encode agent information while lateral regions of anterior lmSTC preferentially encode patient information.

These findings provide preliminary evidence that these sub-regions of lmSTC encode the values of the agent and patient variables. However, it remains open whether and to what extent these sub-regions are specialized for representing agent and patient information—that is, whether they tend to represent one kind of information and not the other. To address this question, we conducted post-hoc analyses that separately defined agent and patient regions within each subject using data from the remaining subjects (See Appendix for detailed description of procedure). Within subjects' independently localized "patient" regions, patient identification accuracy was significantly greater than agent identification accuracy across subjects (lateral lmSTC: t(24)=2.99, *p*=0.006). Within the two "agent" regions, agent identification was above chance (t(24)=2.04, p=0.02: t(24) = 2.38, *p*=0.01), and patient identification was not (t(24)=0.86, *p*=0.20; t(24)=-0.29, *p*=0.39), but the direct comparison of accuracy levels for agent and patient identification was not statistically significant (*p*=0.27 and *p*=0.15).

To further assess the role-specificity of these sub-regions, we localized a large portion of the anterior lmSTC in a manner that was unbiased with respect to its role-preference, and then quantified the average preferences of slices of voxels at each X-coordinate (See Appendix). We found a clear trend in role-preference along the medial-lateral axis, with medial portions preferentially encoding agent information and lateral portions preferentially encoding patient information (Figure 3c).

We attribute the observed effects to differential representation of the values of abstract semantic role variables across this region. One might wonder, however: could these effects simply be due to differences in motion representation between the agent and patient roles? The superior temporal cortex is known to be involved in the representation of biological motion (Blake & Shiffrar, 2007; Han et al., 2013), and our events were all characterized by the movement of one event-participant with respect to another. However, this seems unlikely for a number of reasons. First, it is unclear why a region that represented motion information would carry information about the identity of the patient, who is often motionless in the event (e.g., 'approached the man'). Although there is a difference in the amount of motion between the agent and patient variables, there is no reason to expect a difference between different patients (e.g., "approached the man", "approached the dog", "approached the cat"). Such a difference between patients would be required if the observed effects were due to differences in motion representation, since we localize the patient region by discriminating different values of that variable.

This only pertains to the patient region, however. Might it still apply to the agent regions, in which differences in motion-content between the various agents are more plausible? The posterior superior temporal sulcus is known to be involved in the representation of biological motion (Blake & Shiffrar, 2007) close to the posterior agent ROI we find here. Moreover, Han et al. (2013) recently reported a right hemispheric region that is selective to motion that is caused by a human-agent, relative to other types of biological motion. Although this human-agent motion region is in the contralateral (right) hemisphere, it is in a similar anterior-posterior location in superior temporal cortex as our anterior agent ROI. If the ability to identify the agent specifically reflects differences in the representation of the motion of different agents, rather than general role-filler bindings, then the nature of the motion involved in an event should affect the classifier's ability to generalize to that event. Specifically, it should be easier to generalize role/filler patterns to events with similar manners of motion than it is to generalize these patterns to events with very different manners of motion. However, post-hoc analysis show that the anterior agent region is just as good at identifying the agent of the verb "blocked" in which the agent does not move at all (p=0.013), as it is in identifying the agent of "chased" (p=0.02) (See Table 2). Likewise, it can also generalize to the agent of "scratched", (p=0.005). These events involve different degrees and manners of motion, but the patterns are similar across verbs, suggesting they do not reflect motion representations[5]. It is unclear why a region that encoded motion-content would have the same patterns of activity for the propositions "the man blocked the dog" as "the man chased the cat", which are then different from the propositions "the girl blocked the dog" and "the girl chased the cat". The results are therefore more consistent with the claim that these regions represent the values of abstract semantic role variables.

[5] However, it is also unlikely that the contralateral juxtaposition of our left anterior ROI and that of Han et al's. (2013) is a coincidence. Homotopic regions of the two cerebral hemispheres have long been known to be strongly, and directly, connected through the corpus collosum (Arnold, 1838). The right hemisphere region (rmSTC) would thus be expected to be strongly connected to lmSTC. If the medial lmSTC selectively represents the value of the agent variable, as we suggest, the location of the human-agent motion processing region could make sense: representing aspects of human-agent caused motion requires first knowing who or what the agent is (the value of that variable). The right hemispheric human-agent motion region would be situated in an anatomical location that allows it to easily read the identity of the agent from its contralateral partner, through their direct collosal connections.

Table 2.

(A). Generalization by verb-context

Verb	Patient in Patient ROI	Agent in Anterior Agent ROI	Agent in Posterior Agent ROI
'Chase'	2.12, p=0.022	2.04, p=0.026	1.59, p=0.063
'Block'	1.67, p=0.054	2.37, p=0.013	2.16, p=0.021
'Bump'	3.03, p=0.003	1.78, p=0.044	2.37, p=0.013
'Approach'	2.15, p=0.021	2.61, p=0.008	3.19, p=0.002
'Scratch'	4.31, p=0.0001	2.77, p=0.005	1.87, p=0.037

Table 2. Classification performance in the three ROIs identified by searchlight analysis. The statistics reflect the ability of role-filler patterns to generalize to a given verb, from the remaining set of verbs, within each ROI.

In a separate search analysis, we found no regions of lmSTC to carry information about the surface subject and surface object of the sentence. For example, no lmSTC region encoded "the dog chased the man" and "the dog was chased by the man" as similar to each other, but different from "the man chased the dog" and "the man was chased by the dog". Within lmSTC, the encoding appears, instead, to be of the underlying agent and patient of the sentence, independent of which noun serves as the sentence's surface grammatical subject or object, consistent with Experiment 1. We were also interested in whether classifiers trained to decode the agent or patient *solely* on one surface form can automatically generalize to the alternate form. Our principal analysis trained on data from both active and passive sentences, subsuming them under the same function. To assess this, we (1) trained on active sentences and tested on passives, and (2) trained on passives, and tested on actives. The results of these two procedures were then averaged and pooled across subjects. We found that the neural representations in both the patient ROI and the posterior agent ROI automatically generalize across active and passive sentence forms

($t(24) = 2.10$, $p=0.023$, one-tailed: $t(24)=1.83$, $p=0.039$), but that the anterior agent ROI patterns did not ($t(24)=1.10$, $p=0.14$). While this difference may signal a functional difference between the two agent ROIs, it is important to note that this training procedure uses 50% of the data used by the searchlight analyses (288 trials vs. 144 trials) making it difficult to interpret this null result conclusively.

A final searchlight analysis within lmSTC identified two additional sub-regions supporting identification of a sentence's verb (See Figure 4). Here, we simply trained the classifier to identify the verb of a sentence on 5/6ths of the data, and tested its ability to identify the verb on the remaining 1/6th, iteratively treating each 1/6th as test data, and averaging across the six cross-validation folds. The anterior verb sub-region ($p<0.025$; -61, -15, 2) was adjacent to patient sub-region. The posterior verb sub-region ($p<0.0001$; -55, -49, 5) in the posterior STS partially overlapped with the posterior agent region. Post-hoc analyses found that these results were not driven by information about a subset of the verbs (See *Appendix*).

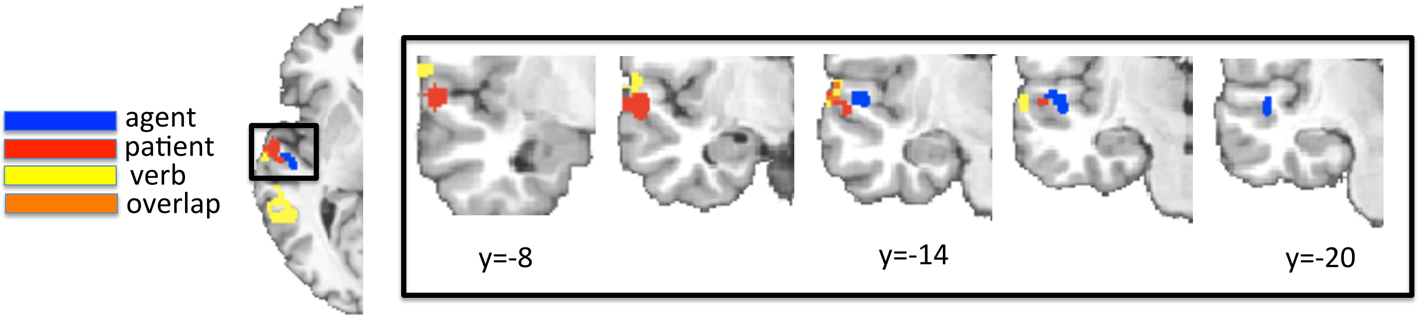

Figure 4. Searchlight results for Experiment 2 for the three classification problems. Coronal slices show the topography of the neighboring anterior verb, agent, and patient sub-regions.

The foregoing analyses strongly suggest that a lateral sub-region of anterior lmSTC selectively encodes information about the identity of the current patient, and somewhat less strongly, that a medial portion of anterior lmSTC selectively encodes information about the identity of the current agent. Together, these results indicate that distinct sub-regions of lmSTC separately and dynamically represent the semantic information sufficient to compose complex representations involving an agent, a patient, and an event-type.

The two experiments presented thus far begin to address an important unanswered question in cognitive neuroscience: How does the brain flexibly compose structured thoughts out of simpler ideas? We provide preliminary evidence for a longstanding theoretical conjecture of cognitive science: that the brain, on some level, functions like a classical computer, representing structured semantic combinations by explicitly encoding the values of abstract variables (Fodor & Pylyshyn , 1988; Marcus, 2001). Moreover, we find evidence that the "agent" and "patient" variables are topographically represented across the upper bank of the left STS and lateral STG, such that adjacent cortical regions are differentially involved in encoding the identity of the agent and patient. At a high-level, these regions may be thought of as functioning like the data registers of a computer, in which time-varying activity patterns temporarily represent the current values of these variables. This functional architecture could support the composition of sentence meaning involving an agent and a patient, as these representations can be simultaneously instantiated in adjacent regions to form complex representations with explicit, constituent structure. These structured representations may in turn be read by other neural systems that enable reasoning, decision-making, and other high-level cognitive functions.

Experiment 3: What variables does left-mid superior temporal cortex represent, and what coding scheme does it use?

Experiment 2 provides evidence that lmSTC encodes the values of abstract variables, consistent with classical symbolic models of cognitive architecture. However, we do not know the precise content of these variables, or the code by which their values are represented. In experiment 3 we address two key aspects of lmSTC's representational content.

First, we ask what coding scheme lmSTC uses to represent a variable's value. Experiment 2 demonstrates that there are reliable patterns of activity for particular concepts in the agent or patient roles. But why does 'dog' have the pattern of activity it does, as distinct from 'man', and 'cat'? 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., 1999) 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. Based on this literature, it is possible that lmSTC could represent words, and do so using a phonological code that treated similar sounding words as similar. Or, it could represent noun-meanings, using a semantic code that treated similar objects as similar[6]. Failure to support either of these models leaves open an intriguing third possibility: that lmSTC does not use similarity-based coding schemes, but instead uses an effectively arbitrary coding scheme. An apparently arbitrary coding scheme maximizes symbol-distinctiveness, and is a sign of an efficient, compressed code (Hopfield, 1982).

[6] It is also possible that it could represent words using a non-phonological code, and concepts using a non-similarity based code. Evidence for either representational format simultaneously provides evidence for the content of the representation. A null effect provides no evidence for the content of the representation.

Second, we ask whether these agent and patient regions are general-purpose slots for encoding “who did it?” and to whom was it done?” across different semantic groups of verbs, or whether these regions specifically represent the relations of participants in the motion-based events of in Experiment 2. In addition to these motion verbs, participants read a set of psychological verbs (e.g., “noticed”), whose core meaning conveys information about the event participants’ mental states. First, we ask whether the agent and patient regions also represent “who did it?” and “to whom was it done?” for these psych verbs[7]. Moreover, psych verbs are unique in that they can vary in whether the entity in the experiencer role is the subject (e.g., “noticed”) or the direct object (e.g., “surprised”) of the sentence. Although it is not guaranteed that lmSTC will carry any information about these psych verbs, if it does, we can ask whether it groups event participants based on their grammatical role as subject or object, or their semantic role as experiencer or stimulus. The motion verbs do not allow for this dissociation, as the agent is always the underlying subject, and patient the underlying object of the sentence. Jointly, asking (1) what coding scheme the agent and patient regions use and (2) how general the regions we have called “agent” and “patient” are will promote our understanding of lmSTC’s representational content.

To evaluate these possibilities, we scanned 40 participants while they read English sentences. Sentences were constructed from a menu of 6 nouns and 8 transitive verbs, to create every possible aRb proposition, excluding propositions in which the same noun

[7] The phrases “who did it?” and “to whom was it done?”, which we have used throughout, are somewhat unnatural when applied to psych verbs, as the verbs’ core meanings do not denote actions. In the sentence ‘A noticed B’, nothing was *done,* narrowly construed. Rather, something just *happened.* We discuss related issues in the final set of analyses for Experiment 3 below. When used here, we simply mean (e.g.,) “who X’ed?”, where X is the verb, and the answer to the question would be realized as the subject of an active sentence describing the event.

occupied both roles (e.g., “the goose approached the goose”). Each participant thus read 6x8x5=240 unique propositions, each presented once. This differs from Experiments 1 and 2, in which each proposition was presented multiple times. Half of the sentences were in the active voice, and half were in the passive voice.

These propositions were constructed from 6 monosyllabic English nouns that refer to animals (‘moose’, ‘cow’, ‘hog’, ‘crow’, ‘goose’, ‘hawk). These nouns were chosen because their phonological similarity relationships are distinct from their semantic similarity relationships. Semantically, these nouns most 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 (‘hawk’ in the bird class). We were guided by the PSIMETRICA model of phonological similarity (Mueller et al. 2003) in determining the phonological similarity of the nouns. To verify the relationships amongst our choices, we also obtained ratings of phonological and semantic similarity from an independent sample of subjects. See the Appendix for more information on stimuli and norming.

Four of the verbs conveyed some aspect of a mental state, which we will, following others, refer to as “psych verbs”, and four verbs did not explicitly convey mental state information, but were characterized partially by different patterns of movement of the agent with respect to the patient in the event. These were ’bumped’, ‘approached’, ‘passed’, and ‘attacked’. We chose four psych verbs whose two semantic roles were labeled ‘experiencer’ and ‘stimulus’ in the VerbNet database (Schuler, 2005). The four psych verbs naturally grouped into two pairs, the members of which were similar both semantically and syntactically: (‘surprised’, ’frightened’) in which the experiencer is the object, and

(’noticed’, ‘detected’) in which the experiencer is the subject. The experiencer-object verbs refer to psychological states caused by the stimulus. The experiencer-subject verbs refer to changes in perceptual state, without explicitly specifying causality. To the extent that there are differences between the two groups of psych verbs, we cannot say whether those differences are syntactic or semantic. However, we are principally interested in what is shared between the experiencer-subject and experiencer-object verbs: that one event participant has a particular type of experience, and the other participant is the cause or object of that experience.

Replication of Experiment 2. We first sought to replicate the results of Experiment 2. To do so, we focused exclusively on the data generated by the 4 motion verbs. There are a number of important differences between this experiment and Experiment 2 that make a replication non-trivial. First, experiment 3 uses only nouns that refer to animals, and not nouns that refer to humans. Bilateral areas of the superior temporal gyrus are engaged when subjects retrieve conceptual knowledge about humans, relative to animals (Zahn et al., 2007). And we noted above that a right-hemispheric region in a similar anterior-posterior location specifically represents human-caused motion, relative to other types of biological motion (Han et al., 2013). Although the fact that we localize the region in Experiment 1 using nouns referring to both animate and inanimate objects suggests the current region is not specific to representations of humans, it is important to evaluate this more directly. Second, each proposition was read just once by each participant. Third, Experiment 3 has much less data with which to train the classifier, largely due to the inclusion of the psychological verbs. In Experiment 2, each classifier was trained on 72 instances of each of 4 concepts in each of the agent/patient roles, resulting in 288 total training trials per cross-validation iteration. In experiment 3, there are only 15 instances of each concept, for each of 6 concepts, resulting in 90 total training trials per cross validation iteration.

Using the same training/testing procedure as Experiment 2, we searched the lmSTC ROI for patterns of activity encoding the identity of the agent and patient variables. This entailed training the classifier on data generated by 3 out of 4 verbs, and testing it on data generated by the 4th verb. As in experiment 2, we cycled through the verbs, using each verb as test data, and averaged the performance across cross-validation iterations. We used a liberal voxelwise threshold ($p<0.05$) and corrected for multiple comparisons in a dilated version of the anterior agent and patient regions in experiment 2. We again found a region that carried information about the patient along the lateral bank of the STS/STG in the same general anatomical location as Experiment 2 ($p<0.001$, corrected for multiple comparison). As in Experiment 2, we also found a region that carried information about the value of the agent variable, located medially and posterior to the patient region ($p<0.005$, corrected for multiple comparisons). The topography of the agent/patient regions is very similar to that found in Experiment 2, despite differences in stimuli and voxel size between the two experiments. Figure 5 shows the location of these results in both Experiments 2 and 3. Agent information is once again preferentially represented medially and slightly ventrally to patient information. As in Experiment 2, we found a highly significant region by role interaction ($F(1, 39)=20.94$, $p<0.001$), demonstrating that these regions differ significantly in the information they carry about the values of the agent and patient variables. We therefore replicate the principal finding of Experiment 2 with new subjects, new stimuli, unique propositions that are read only once, less training data, and different scanning parameters.

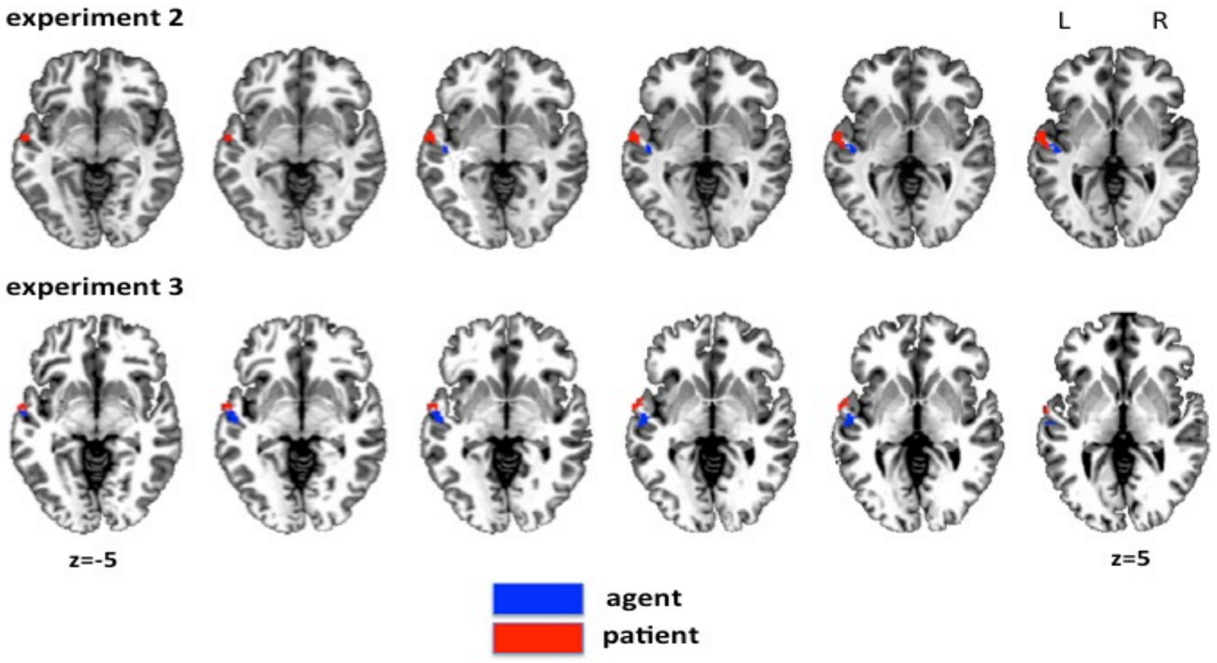

Figure 5. Agent and patient searchlight results for Experiment 2 ($p<0.005$ voxelwise, $p<0.05$ clusterwise corrected) and 3 ($p<0.05$ voxelwise, $p<0.005$ clusterwise corrected). We see a similar medial-lateral topography, whereby medial regions encode the identity of the agent, and lateral regions encode the patient.

Representational Similarity Analyses. We next asked whether the neural similarity of representations in these ROIs correlates with the phonological and semantic similarity models of the nouns. The semantic similarity model is a representation of how semantically similar each pairwise combination of nouns (e.g., “moose/goose”, “moose/cow”, etc.) was rated to be. See Figure 6a. Likewise, the phonological similarity model is a representation of how phonologically similar each pairwise combination was rated to be. For each participant, we correlated the patterns of activity for every possible pairwise combination of nouns creating a neural similarity matrix representing the similarity of the neural patterns for each of the 15 (6*5/2) pairwise noun comparisons. We then carried out the representational similarity analysis by correlating this neural similarity matrix with each of the representational similarity models. This analysis is thus a correlation of correlations, and is called a 'Representational Similarity Analysis' (RSA) (Kriegeskorte et al, 2008). Across subjects, we found that the neither the phonological nor the semantic similarity models predicted neural similarity in the agent region (semantic: t(39)=0.14, *p*=0.88; phonological t(39) = -1.04, *p*=0.3), nor within the patient region, (semantic: t(39) = 0.51, *p*=0.61, phonological t(39)=-1.98, *p*=0.055). Thus, the only comparison that approaches significance is the phonological similarity model, within the patient region. However, this trend actually goes against the direction predicted by the model: the more phonologically dissimilar two words are, the more similar their neural representations tend to be. Given that this effect is not statistically significant on its own, let alone correcting for the four tests performed, this trend should not be taken too seriously. We interpret these as 4 null effects. These null effects may yet be informative, however.

Whether it is a meaningful null result depends partly upon whether there are other regions whose neural similarity structure correlates with the similarity structure of our models. If there exist brain regions that exhibit the similarity structure predicted by the models, then it suggests the models aren't deficient, and the agent and patient regions simply may not use these encoding schemes. To test this, we searched the left superior temporal cortex for regions, perhaps outside the agent and patient regions, whose patterns of activity were predicted by the representational similarity models.

To do so, we first coded trials based on the nouns present on a given trial, without respect to the roles those nouns occupied. This differs from previous analyses, in which we coded for whether a word was present in a particular role. We thus looked for phonological and semantic representations that were role-invariant. However, this coding scheme produces 4 times as much data as the above similarity analysis, given that each word appears as the agent, patient, stimulus, and experiencer an equal number of times. Thus, for each of the six nouns, we randomly selected a subset of 20 trials, yielding 120 total trials for the analysis in order to equate the number of trials to the similarity analysis in our agent and patient ROIs. Different random subsets of trials were selected for each participant.

We searched the lmSTC region used in Experiment 2, which encompassed a large part of superior temporal gyrus, sulcus, and the middle temporal gyrus (See Figure 2a). We found that phonological similarity correlated significantly with neural similarity in the left posterior superior temporal sulcus ((-54, -8, -9), p<0.005, voxelwise, k=18, p<0.05, corrected). See Figure 6. The posterior temporal sulcus is involved in speech processing (Hickok & Poeppel, 2007), and, moreover, its activity is modulated specifically by phonological variation (Chang et al., 2010; Vaden Jr. et al., 2010). For example, it carries information about vowel sounds (Formisano, 2008) and exhibits decreased responses to repeated phonological content in a repetition-suppression paradigm (Vaden, Jr., 2010). The regions reported in these previous studies are very close to the location of the present effect. For example, the center of the region reported in Vaden Jr., et al. (2010) is -54, -12, -12, while the current center is -54, -8, -9, just a few millimeters away. Our analysis therefore appears to be sufficiently powered, and use an adequate model to detect regions that encode phonological similarity relationships. This finding increases our estimate of the likelihood that the agent and patient regions do not encode these relationships.

Next, we performed the same analysis using the semantic similarity model. We searched the left superior temporal cortex for regions whose neural similarity tracked the semantic similarity of the nouns, without respect to the role they played in the sentence. Using the same voxelwise threshold as the phonological similarity search (p<0.005), we found no clusters that withstood correction for multiple comparisons. However, at a lower voxelwise threshold (p<0.01, voxelwise), we found a trending, but non-significant, cluster of interest. This cluster (k=8) was in the anterior left superior temporal sulcus, bordering the middle temporal gyrus (-54, -8, -9). See Figure 6. This cluster, located in the anterior STS, contained the peak voxel in lmSTC for this analysis (p<0.001). This region is of interest because it has previously been found reflect semantic processing of verbal information, over and above phonological and syntactic processing. For example, Vandenberghe et al. (1996) report a nearby region (-40, -12, -10) that is activated for semantic relatedness judgments. This is expected if this region encodes semantic similarity relationships, given that semantic relatedness judgments need access to representations of semantic similarity. In a subsequent study, Vandenberghe et al. (2002) found a nearby region of anterior STS to covary with the meaningfulness of a sentence. Finally, Pallier et al. (2011) report a nearly identical region (-54, -12, -12) in which the magnitude of activity in aSTS parametrically tracks the number of meaningful constituents in a sentence, but not when those constituents are meaningless, but syntactically well formed jabberwocky. These studies all isolate semantic processing, over and above syntactic and phonological processing, and find activity within a few millimeters of one another in aSTS. This suggests that this cluster, though non-significant when correcting for multiple comparisons, may truly reflect the semantic similarity structure of the nouns. However, we cannot establish this conclusively. There may, of course, be other brain regions that encode the semantic relatedness of the stimuli, and perhaps do so more robustly. Past literature suggests these might be in the ventral temporal cortices or the posterior middle temporal gyrus (Kriegeskorte et al., 2007; Connolly et al., 2012; Fairhall & Caramazza, 2013). However, in keeping with the other analyses of Experiments 2 and 3, we restricted our search to the left superior temporal cortex.

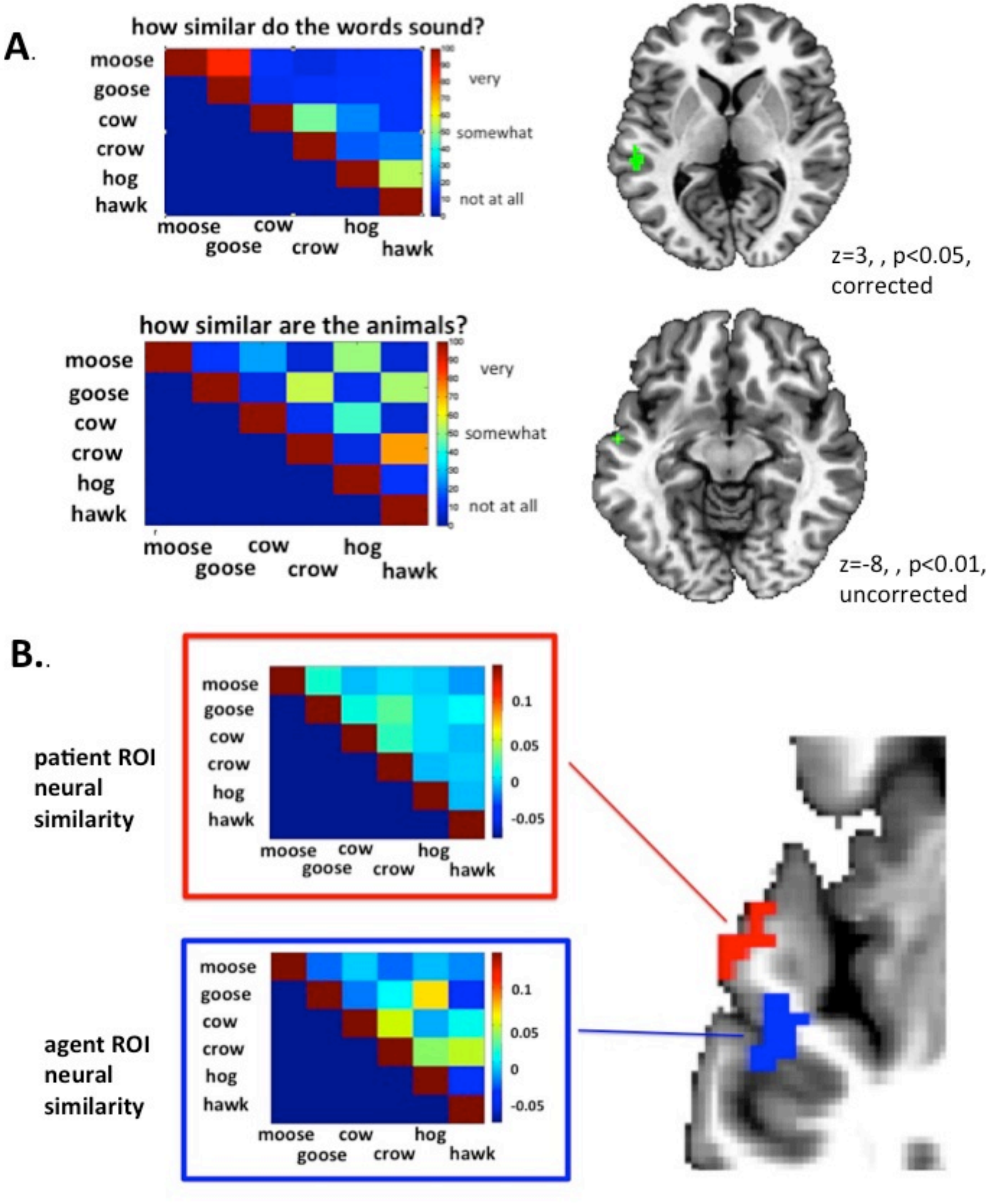

Figure 6. (A). LEFT: Phonological and semantic similarity spaces for the stimuli, derived from participants' pairwise similarity ratings, and RIGHT: regions in lmSTC in which neural similarity correlates with each of these models. The region of posterior STS correlates with the phonological similarity model correcting for multiple comparisons, while the semantic region in aSTS merely trends. (B). The observed neural similarity structure of the agent and patient regions of lmSTC. There is no visually discernible similarity structure in the stimuli used.

These similarity-based search results thus uncover a region of the left superior temporal cortex in which neural similarity structure correlates with the phonological similarity structure of the nouns. There is a hint of a correlation in an appropriate region for the semantic similarity model. But this does not withstand correction for multiple comparisons. These results increase our estimate of the likelihood that the agent and patient regions simply do not encode the phonological similarity of the nouns. It is doubtful that they encode the taxonomic distinction we test here (mammals vs. birds). However, failure to support one particular semantic division (mammals vs. birds) does not mean this region does not encode any semantic distinctions.

How lmSTC encodes the values of particular variables remains unanswered, however. Inspecting the observed neural similarity matrices provides little guidance, as all the correlations are low (See Figure 6). One possibility is that lmSTC uses a non-similarity-based code to represent each value. Such an apparently arbitrary code would allow the values of these variables to be easily distinguished by other neural systems. Moreover, data that has been compressed for efficient storage can also appear arbitrary to an outside observer (Hopfield, 1982). We discuss the possibility of an arbitrary coding scheme in more detail in the general discussion. For now, we can only say that we have no positive evidence that the code is phonological or semantic.

How wide are the slots? Are these regions general slots that represent “who did it?” and “to whom was it done?” across any type of sentence? Or does lmSTC represent fine-grained information about nature of the event described? To evaluate the nature of the variables, we first asked whether the agent and patient regions, as localized using the searchlight analysis above, could also learn classification functions that group the entire set of sentences by their deep subject and deep object. The deep subject and deep object here would correspond to the “actor”/ “undergoer” roles in role and reference grammar (Van Valin, 1997). By “deep subject”, or “actor” we mean the participant that is realized as the subject in active voice constructions and the surface object in passive voice constructions. By the “undergoer” or “deep object”, we mean the participant that is realized as the object in active voice constructions, and the surface subject in passive voice constructions. If these are general slots encoding the actor and undergoer across types of events, we should be able to decode values of these variables from the agent and patient regions for psychological events as well.

To evaluate this possibility, we first treated all sentences containing the same deep subject or object of the sentence as identical. For example, “the moose bumped”, “the moose noticed”, and “the moose surprised” were coded identically (along with their corresponding passive versions). Using this procedure, we found that the agent and patient region could not classify the deep subject (t(39)=-0.58, *p*=0.56) or deep object (t(39) = 1.45, *p*=0.08), respectively, across all sentences. This is despite the fact that this paradigm has twice as much training data as the agent/patient classification problems, and 50% of the data is identical across the narrow (agent/patient) and broad (deep subject/deep object) classifications tasks.

Some theories of the linguistic structure of psych verbs suggest that experiencer-object, but not experiencer-subject, verbs require an additional movement operation, internal to syntax, that relocates the experiencer from the subject to the object position (which may perhaps be reversed in comprehension) (Belletti & Rizzi, 1988; See Baker, 1997 for discussion). If true, one might think that this additional syntactic movement operation, present in 25% of the stimuli, is disrupting the classifier's performance. We thus performed the same analysis as above, but with the experiencer-object verbs removed, and again found that we could not decode the subject, as such, from the agent region (t(39)=0.46, *p*=0.64), nor the object from the patient region (t(39)=-1.57, *p*=0.06, this trend is toward significantly below-chance performance).

We next asked whether any regions, perhaps outside the agent and patient regions in lmSTC, could reliably identify the deep subject or deep object, across motion and psych verbs. This searchlight analysis revealed no regions that represent the deep subject (i.e., "who did it?"), as such, across all verbs. This was true regardless of whether the experiencer-object verbs were included. However, this searchlight analysis did identify a region of the superior temporal gyrus that carried information about the deep object ("to whom was it done?"), (p<0.05, corrected, k=32, center: 47, 15, 1). This region was dorsal and medial to the agent and patient regions, and almost completely non-overlapping with the agent and patient regions. Only one voxel overlaps between the agent and deep object clusters. No voxels overlapped between the patient and deep object clusters. (See Figure 7).

These results indicate that the agent and patient regions are not general slots for representing the underlying subject or object of the sentence. Rather, at this stage, the evidence suggests that they may be more narrowly implicated in encoding who did what to whom in a class of motion events. However, these data provide no positive evidence regarding how the semantic relations of psych verbs are represented. Given that all the theoretical considerations that have motivated experiments 1 and 2 apply to psychological verbs as well, this question is of considerable interest.

Thus, we next asked whether we could find analogous regions in lmSTC that represent the relations of psych verbs. There are a number of possible outcomes. First, experiencer and stimulus could have unique and separable regions, like agent/patient. These regions would not have to be in lmSTC, but, in this study, we focus on the same broad region of the left superior temporal gyrus and sulcus. Second, it's logically possible that experiencer and stimulus could be mapped on to the agent and patient regions. Even though the agent and patient regions do not encode the deep subject or object for psych verbs, they could still represent the experiencer and stimulus of these stimuli, as these are separate functions. It is unclear, however, how one would expect this mapping to proceed, as the experiencer has both agent-like (mental state) and patient-like (affected, state-changed) properties (Dowty, 1991). Nonetheless, it is possible that these semantic roles could map directly to the agent/patient regions. Finally, lmSTC could carry no information about the roles of the psych verbs, which would further suggest that it is narrowly concerned with the relations of participants in a class of motion events. On the one hand, there is considerable cortical specialization in conceptual domains of object knowledge (Mahon & Caramazza, 2009). It is possible that there could likewise be cortical specialization for encoding event-types. On the other hand, the domain-general nature of compositionality suggests that there is some explanation for why the same principles apply across different types of events. This does not, however, require that all structured semantic content is represented by the same cortical substrate.

To evaluate these possibilities, we first simply searched the lmSTC ROI, formed by dilating the anterior agent/patient regions of Experiment 2 by 8 mm, for neighborhoods that could identify the experiencer and stimulus, as such, across syntactic configurations.

This analysis was structurally identical to the agent and patient searchlight analyses. But instead of using the data generated by the motion verbs, we used only the data generated by the psych verbs. We separately trained the classifiers to identity the experiencer or stimulus on 3 out of 4 verbs, and tested them on the held-out verb. The experiencer classification function would, for example, group "the moose noticed" with "surprised the moose", despite the fact that 'moose' occupies different syntactic positions in these sentences. In both cases, the verb's core meaning conveys information about the moose's mental state. Here, we found one anterior region that carried information about stimulus (*p*<0.05 corrected, k=40, center= 49, -3, -2), and one lateral region that carried information about the experiencer (*p*<0.05 corrected, k=45, center=60, 8, -3). The stimulus sub-region is anterior to the agent and patient regions, occupying some of the medial portion of the superior temporal gyrus. Post-hoc analysis revealed that this region could not successfully classify the experiencer (t(39)=1.16, *p*=0.13), but also was not significantly better at identifying the stimulus than identifying the experiencer (t(39)=0.88, *p*=0.38). Its selectivity is therefore questionable. The experiencer region is nestled between the agent and patient sub-regions, with some voxels overlapping each of the agent and patient sub-regions. See Figure 7. Despite this overlap, we find that this ROI is unable to significantly decode the agent (*p*=0.21) or patient (*p*=0.15) for the motion verbs. It demonstrates a non-significant trend toward identifying the stimulus (t(39)=1.53, *p*=0.067). However, in a direct comparison, it is marginally better at identifying the experiencer than the stimulus (t(39)=2.00, *p*=0.052). The interaction between the experiencer and stimulus regions trends, but is not statistically significant F(1,39) = 2.89, *p*=0.09. Figure 7 shows these search results alongside the agent and patient sub-regions.

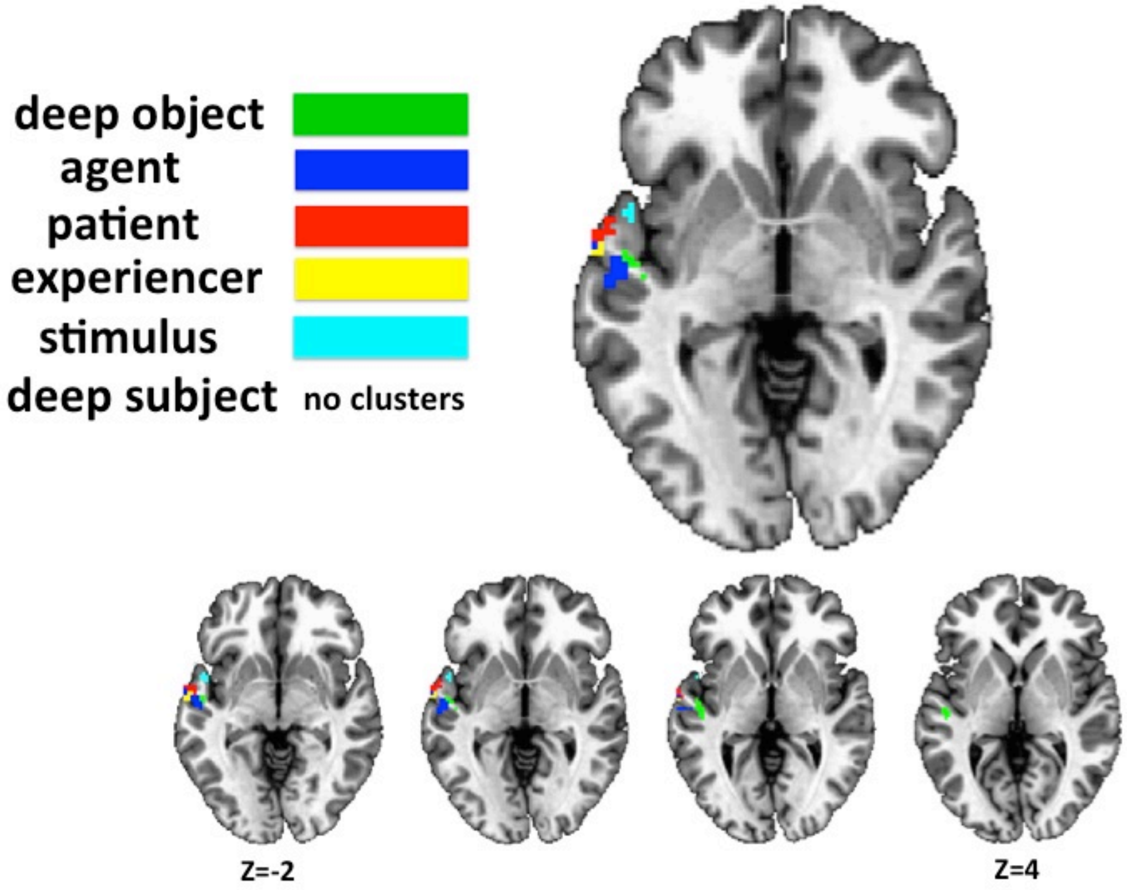

Figure 7. Searchlight analyses within lmSTC for different semantic and syntactic functions (p<0.05, small volume corrected). Z=0. Deep object extends dorsally, and experiencer ventrally beyond what is shown.

We next performed an even stronger test to identify representations of these semantic roles. We trained pattern classifiers to identify the stimulus and experiencer on experiencer-subject verbs (e.g., noticed), and tested them on experiencer-object verbs (e.g., surprised), and vice versa. This requires the classifier to not only group active and passive versions of the same deep syntactic structure together, but to generalize representations of fillers automatically across deep syntactic structures, grouping instead by the role the participant plays in the event. This is a stronger test than the previous analysis, but one that should produce convergent results if these regions encode abstract information about participants' roles in the event.

Here, we found one sub-region of lmSTC that reliably contained information about the stimulus (*p*<0.01 corrected, k=59, center 49, -5, -6). This region overlaps substantially with the stimulus region reported above. Again, it is anterior to the agent and patient regions, and encompassed some of the medial part of the gyrus, internal to the middle longitudinal fasciculus. This cluster is larger than the region identified in the previous stimulus/experiencer analysis, subsuming the previous stimulus region. This is surprising, given that requiring generalization to new syntactic contexts is a stronger test of generalization, and uses less data. This searchlight analysis recovered no sub-regions regions within lmSTC that could generalize across syntactic groups to identify the experiencer. One the one hand, this might suggest that 'experiencer' in verbs of perception is actually a different semantic role than 'experiencer' in verbs of caused-mental state, as some have suggested (Pestsky, 1989; Dowty; 1991; Baker, 1997). If, however, we directly test the experiencer ROI localized by the previous searchlight analysis (which grouped particular role/filler patterns across experiencer-subject and experiencer-object, but did not require generalizing role/filler patterns across them without further training), we find that it can automatically generalize across experiencer-object and experiencer-subject syntactic contexts (t(39)=2.59, *p*<0.007). This suggests that there is some representation held constant for the experiencer in experiencer-subject and experiencer-object constructions that can facilitate generalization of role/filler patterns across the two classes. As expected, a targeted test of the stimulus ROI localized above can also automatically generalize across syntactic contexts to identify the stimulus (t(39)=2.58, *p*<0.007). Taken together, these two groups of searchlight analyses suggest that an anterior region of lmSTC may represent the identity of the stimulus, and a posterior and lateral region may represent the identity of the experiencer. However, we also find that these sub-regions are not as differentially selective as the agent/patient regions, given that we did not find a significant role by region interaction. These results should therefore be taken as suggestive, but perhaps not definitive, evidence that there are separate regions in lmSTC that represent the experiencer and stimulus roles.

We explore one final set of analyses in Experiment 3. We previously tested the hypothesis that lmSTC represents broad roles, grouping narrow roles into the "actor" (or subject) and "undergoer" (object) roles, and found no evidence for this idea. However, these regions could, in principle, still represent alternative, coarsely individuated roles that do not correspond to the subject/object of the sentence. For example, it's possible that the agent regions could encode the function [agent or experiencer], where experiencer is independent of subject/object position, and the patient region could encode [patient or stimulus]. Or, alternatively, the patient region could encode [patient or experiencer], and the agent region [agent or stimulus]. These groupings could derive from the underlying semantic features of the events, shared across roles (Rozwadowska, 1988; Dowty, 1991). Perhaps the agent and experiencer are grouped together based upon information about their mental state. Or perhaps agent and stimulus are grouped together in virtue of causing the event. Although the anatomical differentiation above, in which stimulus region is located anterior to the agent/patient regions may suggest there is no such overlap, this analysis uses twice as much data, and therefore could detect effects hidden from the previous analysis.

We first performed this analysis by simply treating the fillers of each broad, two-role grouping identically. For example, to search for regions that treated the agent and experiencer equivalently, we would code ‘moose’ as agent, and ‘moose’ as experiencer identically and distinct from ‘goose’ as agent and ‘goose’ as experiencer, etc. These classifiers identified the value (Moose? Goose? Hawk? Hog? Cow? Crow?) of these broad role variables for each trial that contained the held-out verbs. This entailed training on 7/8 verbs, testing on the 8th, iterating this process holding each verb out as test data, and averaging results across the eight cross-validation folds. We did so for all four inclusive functions ([agent or experiencer], [agent or stimulus], [patient or experiencer], [patient or stimulus]. This analysis targets regions in which the patterns of activity for particular role/filler pairs generalize across narrow semantic roles, but in systematic ways. We found a statistically significant cluster of voxels in medial lmSTC that grouped the agent and stimulus together (p<0.01, corrected, k=48; center=49, 9, 0) and a statistically significant lateral region that grouped the patient and experiencer together (p<0.001, corrected; k=83; center=57, 3, -6). See Figure 8. This medial agent/stimulus region partially overlapped both the agent and stimulus regions, localized previously. Likewise, for the patient/experiencer region. No regions of lmSTC significantly grouped either the agent with the experiencer or the patient with the stimulus. Decoding performance differed significantly between sub-regions, across roles ($F_{regionXrole}(1,39)=8.17$, p=0.006)), demonstrating that they are dissociable in their informational content.

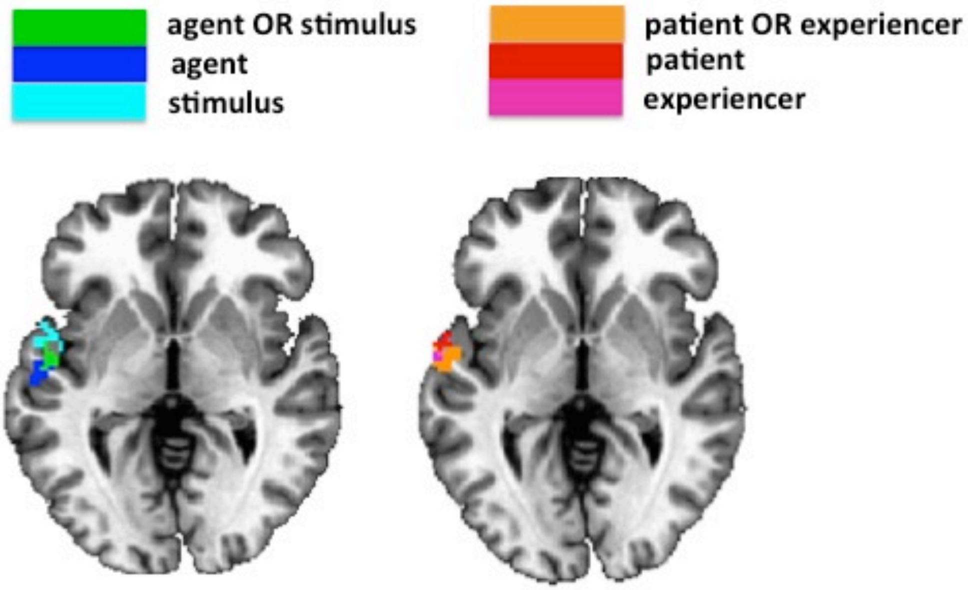

Figure 8. Spatial relationship between regions found to carry information about basic roles, and regions that can successfully identify the value of alternative groupings of the role variables. The experiencer region extends ventrally beyond what is shown here.

We next asked whether patterns learned for fillers in one role (e.g., ‘moose’ as agent) automatically generalize to the other role (e.g., ‘moose’ as stimulus) without further training. For example, we asked whether a classifier trained only to identify the agent can thereby be exported to identify the stimulus, and vice versa. Indeed, the same general portion of the medial STG that subsumed agent and stimulus under one category, medial to the middle longitudinal fasciculus white matter pathway, can automatically generalize from agent to stimulus, and vice versa (p<0.05, k=37, center=47, 10, 1). Likewise, a portion of the lateral STG can reliably generalize from the patient to the experiencer, with no further training (p<0.005, k=60, center=56, 11, 3). See Figure 9. This is a strong demonstration that the same patterns of neural activity are instantiated in a region of medial STG for different concepts when the agent or stimulus. And the same patterns of activity are instantiated in lateral STG when a concept is the patient or experiencer. Within the patient/experiencer region, we find no information about the agent/stimulus (t(39)=-0.39, *p*=0.35). Likewise, the agent/stimulus region can not reliably decode the identity of the patient/experiencer across subjects, though there is a trend in this direction (t(39)=1.48, *p*=0.075). We find no evidence of such generalization between patient and stimulus, nor between agent and experiencer with these verbs. The topography of these regions is related to, but distinct, from the narrow agent, patient, stimulus, and experiencer regions. See Figure 8. These broad-role functions are represented more centrally in the STG, while the narrower functions are represented on the more extreme portions of the gyrus, and in STS.

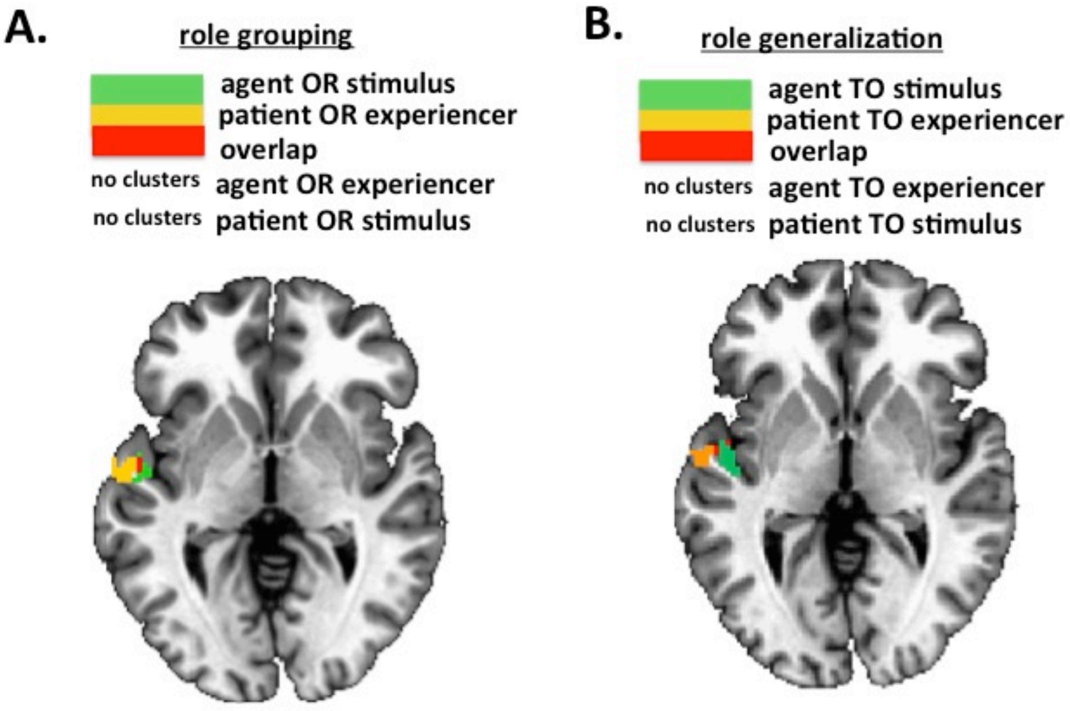

Figure 9. Search within lmSTC for regions that can (A) learn a broad classification function for role-filler patterns (e.g., ('moose' as agent) that subsumes particular narrow roles and (b) generalize decoding of role-filler patterns from one role to another without further training. These are related tests, but (b) provides a stronger test of generalization.

But are these role generalizations meaningful? One possibility is that this medial/lateral topography reflects the 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. However, despite our use of the terms "agent/patient" not all of the verbs are causal agents and patients, in a strict sense. That is, in these verbs, the participant expressed as the first argument does not necessarily bring about a change of state in the participant expressed as the second argument (e.g, "approached"). Nonetheless, there is still a broader sense in which the role we have called 'agent' is the one responsible for the event's happening (approached/ bumped/ passed /attacked), even if they do not bring about a change in the patient. This fact also holds for the stimulus in experiencer-object verbs, in which the stimulus is responsible for the change in psychological state of the experiencer (surprised/frightened). It is less obvious that this analysis works 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. To the extent that they're causal, it's unclear which direction causality runs. The sentence "the moose noticed the crow", could, of course, be interpreted such that the crow has done something that caused the moose to notice it. This, however, does not seem to be part of the core meaning of the verb 'noticed', but an inference that depends upon the particular entities in the relationship.

If this general medial-lateral topography reflects the causal structure of the events, one might expect the classifier's generalization from agent/stimulus to be significantly better for those verbs in which the stimulus is clearly causal. Here, these are the experiencer-object verbs (e.g., ‘surprised’). However, the patterns of activity in this agent/stimulus region generalize from the agent to the stimulus in experiencer-subject verbs (e.g, from “the moose bumped” to “noticed the moose”) (t(39)=2.37, *p*=0.011) just as well as they generalize to the stimulus in experiencer-object verbs (e.g. “the moose bumped” to “the moose surprised”) (t(39)=1.74, *p*=0.045). Although these considerations are not definitive, they suggest that the topography may not reflect causal relationships, in a strict sense. There may, however, still be a more a general notion of causation that is applicable: one in which there is some asymmetric dependence of the relationship described amongst 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 cow not had some pre-existing note-worthy feature. Or in ‘the hawk approached the moose’, the moose would not have been approached, had the hawk not approached it.

A related possibility is that this medial/lateral organization reflects the temporal, but not the causal, structure of the events. Although the causal and temporal structure of an event are highly correlated with one another, temporal precedence relations are a wider class; at least when causal relations are construed narrowly. 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 like ‘noticed’ that aren’t clearly causal. For, the state or action of the stimulus pre-dated the *noticing* event, just as the movement of the agent pre-dated, or instigated, the *passing, bumping, attacking, and approaching* event. Although this idea is clearly distinct from a strict sense of causality in which the agent directly brings about some change in the patient, it is less distinct from the more general notion of causality mentioned above, in which the property/state of the experiencer or patient that is denoted by the verb would not exist without the property/state of the stimulus or agent.

A final 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. It is unlikely that the moose surprised the hawk by its cavalier attitude toward climate change. This idea finds a theoretical basis in Jackendoff's (1990) 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. Moreover, he holds that there is no principled distinction between the semantic representation of a verb's meaning, and the conceptual representation of an event (Jackendoff, 1983; Jackendoff, 1990). This conflation makes it plausible that [GO(thing, place)] could be re-used in interpreting the psych verbs here as well, even if one wouldn't 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 GO, the mover, and the patient/experiencer region might represent the second argument of the GO function, which is the 'place' to which the entity moves. Ontologically, the event participants are *things*, not *places*, but perhaps the content of the place could be 'near *thing*'.

If this movement-based explanation is true, we might find that the patterns of activity in the agent/stimulus region correlate with the semantic similarity model of the nouns: mammals and birds have different types of movement, and if these patterns of activity are specifically representing the movement of the agent/stimulus, this might be captured in its correlation with the semantic similarity model. We find no correlations within the agent/stimulus region with the semantic similarity model, however. This is certainly not decisive, given that these semantic functions need not encode the values of their arguments using similarity-based codes.

It is also worth remembering the post-hoc results of Experiment 2. There, we found the medial agent ROI to be invariant to the manner of motion of the event, generalizing as well to the agent of 'blocked' as to 'chased' and 'approached' (See Experiment 2 above. Table 2). Success on 'blocked' is more consistent with a quasi-causal explanation for the medial-lateral topography than either the motion or temporal explanations: the agent moves little to none, and the patient is moving or intending to move prior to the blocking relation, reversing the temporal order. No verbs in Experiment 3 have this property. However, it is important to remember that the agent and patient ROIs are not anatomically identical to the agent/stimulus and patient/experiencer regions identified in Experiment 3, despite having the same medial/lateral topography. The agent/stimulus region of Experiment 3 is anterior to the agent region of both Experiments 2 and 3, falling medial to the white matter in the STG. Given that we see fine differentiations in the representational profiles of adjacent sub-regions, we should be wary of generalizing from the properties of the agent ROI in Experiment 2 to, the agent/stimulus ROI of Experiment 3.

One possibility is that causal relations are represented posterior to the white matter (explaining the patterning of "blocked" with "chased", "scratched" etc.), while motion relations are represented anterior and medial to the white matter (explaining the patterning of the agent of 'approached', 'bumped', etc., with the stimulus of 'noticed'). (See again Figure 7 for the relationship between the agent and agent/stimulus regions of Experiment 3).

Unfortunately, the present experiment was not designed to differentiate these subtly different explanations for the organization of lmSTC. It will be important for future research to tightly manipulate the presence of motion, as well as the causal and temporal structure of events in order to better understand lmSTC's topographic organization. In future studies, the causation and motion hypotheses could be fractionated by simply including events with more abstract causal relations. For example: "John's attitude surprised Cheryl". Here, there's no literal motion of an entity that could be the causal force, even in a general sense. Differentiating the temporal order of an event's sub-parts from the causal structure of those sub-parts is more challenging, though verbs like 'blocked', 'obstructed', and 'tripped' could prove useful here[8].

Taken as a whole, these analyses reveal a number of interesting properties of lmSTC. First, we find no evidence that the agent and patient regions are general slots for encoding the underlying subject or object, (or "actor" or "undergoer") of a sentence. This series of searchlight analyses suggest a richer, and more complicated, representational topography. The more extreme anterior, posterior, and lateral portions of the region represent aspects

[8] If lmSTC represents events in general and not verb meaning more narrowly, then separating causation and temporal relations becomes easier. One could then include complex events composed of a sequence of sub-events, some of which are causally related and some are not. For example, "The basement of the Jones' house flooded. For days, they removed the water. The foundation was ruined and the house collapsed". Here, there is one event that causes two subsequent events. The temporally intermediate event, their removing water with buckets, does not cause the third event, the collapse of the house. The temporal unfolding of events is thus partly dissociated from the causal relations among events. A causes both B and C. But the temporal ordering is A, then B, then C. Causal and temporal explanations of the medial-lateral topography would therefore predict a subtle difference in the medial-lateral encoding B and C. Causally, they both owe equally to A. But temporally, B precedes C. Of course, this scenario is much more complex than the two-place predicates we have used in these studies. This is well outside the meaning of a single verb. It's unclear whether lmSTC could simultaneously encode multiple events, or how it would do so. We again address how multiple propositions could be represented in lmSTC in the general discussion.

of the event that are not shared across the roles we use here, while the more intermediate regions carry information that generalizes across roles (See Figure 6), but in a systematic way. One could imagine a number of reasons for this general topography. First, it is possible that there is some systematic functional differentiation between narrow and broad roles. It is possible that the regions around the extremity first extract narrow semantic role information that differs across different types of verbs, and the broad-role regions interpret this information to represent aspects of the event itself. This latter interpretation may consist in the application of simple semantic functions such as [GO [thing,place] [STAY [thing, place]] [CAUSE [thing/event, event]] that allow for the observed generalization across types of verbs. It is also possible that the reverse it true; the broad-role regions first map from the syntax of the sentence to coarse categories that are then differentiated to encode aspects unique to these classes of events, differentiating psych from motion events moving outward from the white matter. Finally, it's possible that there is no narrow/broad role differentiation, per se. Rather, these could all reflect semantic functions at the same level of representation. The observed organization could simply reflect which semantic functions happen to recur across the particular verbs we use.

There, are, however, a number of important caveats regarding the results of Experiment 3. First, we have performed a large number of search analyses, and only corrected for multiple comparisons within analyses, not for the number of search analyses we performed in total. We believe this to be a worthy sacrifice in order to fully explore the representational topography of the region. However, some of these effects would not survive correction for multiple searchlight analyses. The basic experiencer and stimulus searchlight results are particularly questionable, given that we do not see a significant interaction between region and role-representation in a direct test. Second, there are statistical dependencies between searchlight analyses that are important to bear in mind when interpreting the results. For example, knowing that the lateral region encodes information about the patient increases the likelihood that it can learn a function that groups the patient and experiencer together. It certainly does not guarantee it, however, as we have seen that a region that can decode (e.g.,) the patient, cannot necessarily also decode the object, even though the former is a subset of the latter. It will thus be important to replicate these results in an independent experiment, as well as to tightly manipulate semantic components of the event to better understand the principles governing the region's apparently nuanced topography.

General Discussion

The sentence comprehension literature and lmSTC

These three experiments provide important information about how the human brain represents a class of sentence meanings. Experiment 1 demonstrates that a broad region of left-mid superior temporal cortex (lmSTC) carries information about who did what to whom in an event. Experiment 1 then develops a pattern-based effective connectivity analysis to link these patterns of activity to affective responses in the amygdala, consistent with a model whereby lmSTC enables the comprehension necessary to produce an affective response to a morally salient sentence. Experiment 2 provides insight into how the lmSTC encodes these meanings, namely by representing the values of the agent and patient variables in spatially distinct neural populations. We find the medial portion of lmSTC to carry information about the identity of the agent, and the lateral portion to carry information about the patient. Experiment 3 replicates the agent/patient topography of Experiment 2, and suggests that the semantic relations denoted by psychological verbs may have a related, but distinct, topography within lmSTC. We find regions that appear to narrowly represent basic semantic roles (agent/ patient / stimulus/ experiencer), and anatomically intermediate regions in which the patterns generalize across these roles, but in systematic ways. The code by which lmSTC encodes the values of these variables remains unclear, however. We find no positive evidence that it is phonological or semantic, leaving open the possibility that lmSTC prioritizes distinctiveness and efficiency in its representations.

These results are broadly consistent with previous research concerning the neural loci of structure-dependent semantic processing while, at the same time, offering new insight into how the meaning of simple sentences is represented. The left superior temporal cortex is one of the perisylvian regions routinely implicated in studies of phrase and sentence-level semantic processing using both functional neuroimaging and lesion data (Friederici et al., 2003; Fedorenko et al., 2011; Meltzer et al., 2010; Pallier et al., 2011; Hagoort & Indefrey, 2014; Humphries et al., 2006; Vandeberghe et al., 2002; Dronkers et al., 2004; Wu, Waller, & Chatterjee, 2008). For example, numerous studies report increased activity in this region during the comprehension of meaningful sentences relative to unstructured word lists (Mazoyer et al., 1993; Vandeberghe et al., 2002; Humphries et al., 2006; Fedorenko et al., 2010). However, lmSTC is by no means the only such region, as anterior regions of the temporal lobe (Rogalsky & Hickok, 2009; Humphries et al., 2006; Pallier et al., 2011) left inferior parietal lobe (Humphries et al., 2006; Pallier et al., 2011), and left inferior frontal cortices (Hagoort & Indefrey, 2014) show a similar representational profile. Moreover, it has been unclear what semantic and/or syntactic information lmSTC represents (Grodzinksy & Friederici, 2005; Rogalsky & Hickok, 2009; Friederici et al., 2003; Vigneau et al., 2006). The three studies reported here suggest that lmSTC is more narrowly involved in encoding the values of semantic role variables. This narrower claim is consistent with multiple pieces of pre-existing experimental evidence.

First, Friederici et al., (2003) find that mid-left STG/STS responds more to implausible noun-verb combinations (e.g., "the thunderstorm was ironed") than to syntactic anomalies ("the blouse was on ironed"). This suggests that left mid-superior temporal cortex integrates lexico-semantic information about a verb and its arguments. However, as we noted in the introduction, it is unclear how these error-detection signals relate to the underlying compositional processes themselves (Rogalsky, & Hickok, 2009). So although Friederici et al.'s finding is suggestive of a narrower role in role assignment or some downstream consequences thereof, it is not, in and of itself, strong evidence that lmSTC represents the values of semantic roles variables. Moreover, other work using the same paradigm has found different regions generating the same violation-of-expectation signals (Kuperberg et al., 2003), and not lmSTC.

More directly, Devauchelle et al. (2009) found that the repetition of a sentence's meaning produces adaptation effects in lmSTC. This paradigm exploits the fact that a neuron that codes for some stimulus-value will decrease its firing rate upon repeated presentations of that value. Population-level neural activity, and consequently the BOLD signal decrease accordingly. The authors report that the left superior temporal cortex adapts to repetitions of sentence meaning, even when that meaning is expressed using different surface syntactic forms, such as the active and passive voice. For example, the superior temporal cortex responded less when a subject read the sentence "an architect designed this palace" if he/she had just read the sentence "this palace was designed by an architect", than if he/she had just read "Some scientists invented a medicine". These semantic adaptation effects occur in mid-STG and mid-dorsal MTG/ventral STS when sentences are presented aurally, and in mid-dorsal MTG/mid-ventral STS when presented visually[9]. These regions correspond closely to those we report here. However, Devauchelle et al. did not include a condition that matched basic lexico-semantic content, but varied structured semantic content. As a result, it is unclear whether the adaptation effects owe to the repetition of propositional content, or the repetition of basic lexico-semantic content, regardless of how the constituent concepts were assembled. They did not, for example,

[9] We note also that Devauchelle et al.'s (2009) effects are relatively weak, and do not withstand whole-brain correction for multiple comparison.

show that that this region adapts to (e.g.) "the truck hit the ball" after "the ball was hit by the truck", but not "the ball hit the truck". We thus extend their findings in important ways.

Perhaps the strongest supporting evidence for our claim comes from a neuropsychological study by Wu, Waller, & Chatterjee (2008). These authors report patients with damage to the mid-portion of the superior temporal gyrus, sulcus, and middle temporal gyrus who have specific deficits in determining who did what to whom in response to both sentences and visual scenes representing actions. In this study, patients were presented with simple agent/verb/patient sentences (e.g., "the circle kicks the square"), or pictures depicting the same event. They were asked to perform a number of tasks, including matching the sentence to a picture depicting the sentence denoted, inferring the consequence of the sentence, or inferring the consequence of the picture. Damage to lmSTC impaired patients' performance on these three tasks, but not their ability to judge the semantic similarity of triads of nouns or triads of verbs. The damage was thus restricted to tasks requiring an understanding of who did what to whom in a sentence or picture. This region appears to correspond very closely to the parts of lmSTC in which we find semantic variables to be topographically represented, although the authors do not provide anatomical coordinates.

In introducing Experiment 1, we noted that reversible sentences have had a long and productive history of use in the neuropsychology of language (Caramazza & Zurif, 1976; Schwartz, Saffran, & Marin, 1980; Ansell & Flowers, 1982 Caramazza & Micelli, 1991; Dronkers et al., 2004). Lesion-mapping and functional neuroimaging tend to converge on important roles for the left inferior parietal lobe (Caramazza & Micelli, 1991; Richardson et al., 2010; Meltzer et al., 2010; Thothathiri et al., 2012) and, to a lesser extent, the left inferior frontal cortex (Meltzer et al., 2010; but see Thothathiri et al. 2012) for comprehending reversible relative to nonreversible sentences. (But see Meltzer for one case implicating the STG as well). Given our use of reversible sentences, one might wonder why we did not converge on the same regions. There are, however, important differences between these studies and our own. First, it is important to remember that we did not contrast reversible and non-reversible sentences, but instead looked for different patterns of activity for different reversible sentences. The reversible/non-reversible contrast specifically isolates syntactic processing, robbing one of the use of semantic heuristic processes to assign thematic roles (Caramazza & Zurif, 1976). These syntactic operations (or the mapping from syntax to thematic roles) isolated by the reversible/non-reversible contrast were present in all the sentences we used, and equally so. Thus, what differs between reversible and non-reversible sentences was held constant across the propositions we classified in Experiment 1, so we would not necessarily expect these approaches to converge on the same neural substrates.

However, given that these syntactic computations are necessary precursors to representing the meaning of the sentence (e.g., analyzing word-order to produce the appropriate syntactic tree), we would expect these parietal or frontal regions to be anatomically and functionally connected to, and feed lmSTC. Indeed, Turken & Dronkers (2011) find the mid-anterior STG to be anatomically connected to the inferior frontal cortex through the uncinate fasciculus pathway, and to the inferior parietal lobe through the middle longitudinal fasciculus. They analyze the functional correlations of these brain regions during rest, and find that mid-anterior STG co-varies with the left inferior frontal cortex and the extent of the superior temporal gyrus, but little with the parietal lobe.

However, it's unlikely that resting state functional connectivity is a good proxy for the computational pathway sub-serving the comprehension of reversible sentences; parsing the syntax of a sentence and using that syntax to assign thematic roles is a paradigmatic analytic process, and a blank screen and the hum of a scanner during rest provide little to analyze syntactically.

But, one might wonder, shouldn't our pattern classifiers have *also* decoded these syntactic computations, since these also differ systematically between "the truck hit the ball" and "the ball hit the truck"? The particular computations executed differ between active and passive sentences, because the function that maps from word order to meaning is reversed. A region that takes, as input, information about (e.g.) word order and produces a representation of the sentence's syntax should have difficulty learning a classification function that grouped "the truck hit the ball" and "the ball was hit by the truck" together. By analogy, a classifier would not treat the computations, as implemented in a physical system, used in the expression '2X4 = 8' as equivalent to '4x2 = 8'. Only the representation of the output is identical.

Instead, we use reversible sentences as experimental tools because they are matched in basic lexico-semantic content, syntactic tree structure, and summed-word frequency. They are thus useful in isolating higher-order semantic representations. But though we use only reversible sentences, our working model is not that lmSTC only encodes the meaning of reversible sentences, but also non-reversible sentences that require binding values to semantic variables. We do not test this here, and this assumption may be wrong. There is limited evidence on whether this region is involved, more generally, in other cases of variable/value binding. What evidence exists is equivocal, but highly relevant.

In the introduction, we noted a study by Dronkers et al. (2004) that detailed the relationship between damage to various left perisylvian regions, and comprehension of different sentences types. For patients with damage to the mid-anterior superior temporal cortex, this comprehension impairment is most severe when the sentence contains some "ambiguity", as Dronkers et al. say, in role assignment[10]. Of all the construction types Dronkers et al. tested, only the comprehension of sentences describing possession relations (e.g., "the girl has a baseball") was spared upon lmSTC damage. Patients were moderately impaired in interpreting simple declaratives (e.g, "the girl is sitting"), though these difficulties were not as marked as those involving verbs with two or more arguments. Notably, possession and the simple declarative were the only construction types that did not denote relations between two or more event participants, each of whom could plausibly occupy multiple roles in the sentence. Patients had marked deficits on all nine sentence types that had required the assignment of multiple roles.

First, we should pause to note that these results are highly consistent with our own; both implicate a region of the mid-anterior superior temporal cortex in the comprehension of two-place predicates. However, despite their general consistency, Dronkers et al.'s results also suggest important limits to the present findings. Given that possession was spared upon lmSTC damage, it seems lmSTC is not involved in all types of semantic composition. Whether this is due to a difference in the semantic content of the relation, the number of arguments, or the particular composition algorithms applied is unclear.

[10] Contrary to Dronkers et al.'s use of 'ambiguity', there's little actual ambiguity in assigning thematic roles based on the syntax of a reversible sentence; only an increased demand on syntactic analysis of the sentence, per Caramazza & Zurif (1976).

Dronkers et al. favor the multiple arguments explanation, suggesting that this region is involved in resolving ambiguities in role assignment, such as those encountered in reversible sentences. However, it is important to note that their patients with STG damage were also significantly impaired on declarative sentences denoting states ("the girl is sitting"), just less so than with verbs of two or more arguments. On the one hand, the fact that patients' comprehension of constructions with multiple arguments is worse than the their comprehension of verbs with one argument might suggest that lmSTC is particularly important for representing two-place predicates, or even more narrowly, to reversible sentences. But on the other, these patients are also significantly impaired on simple one-place predicates denoting states, like "the girl is sitting", relative to controls, although Dronkers et al. minimize this fact in their interpretation.

The Wu, Waller, & Chatterjee (2008) study described above also contains data that bear on the scope of the relations represented by lmSTC, but instead of the number of arguments, their data bear on the types of relations encoded. In addition to tests of semantic role assignment in patients described above, the authors included sentences describing spatial relations ("the circle is above the square") in their battery. They found that difficulty with spatial relations is largely predicted by damage to the left parietal lobe, in contrast to difficulty with thematic relations, which is predicted by damage to the left temporal region of present interest. The authors lean on this difference in their interpretation of the results, contending that these are two separate neural systems, whose relative anatomical location depends on their proximity to the spatial representations in the parietal lobe, and motion representation in the temporal lobe, respectively. However, they also find a small part of lmSTC in which damage significantly predicts difficulty with spatial relations, though this relationship is generally neglected in their interpretation. This region of STG is small relative to the region that predicts difficulty with thematic relations, and also relative to the region of the parietal lobe sub-serving comprehension of spatial relations. Nonetheless, damage therein significantly predicts impairments in the comprehension of spatial relations. Anatomically, this region falls near the sub-region in which we find information about the stimulus-role to be represented.

A recent study by Miozzo et al. (2014) is also relevant here. The authors found two patients with left frontal lesions that had difficulty understanding relations conveyed by spatial prepositions ("the circle is on the square") and adjectival comparisons ("the girl is taller than the boy"), but not relations communicated by verbs ("the woman helps the man"). This suggests that neural and cognitive systems for representing different types of reversible relationships are dissociable. However, it does not show that they are doubly dissociable, and hence that the converse pattern would hold. That is, damage to lmSTC could, in principle, impair comprehension of spatial and adjectival relations, even if frontal damage spares role assignment for verbs. This asymmetry is particularly plausible given that Miozzo et al., contend that non-verbal lexical categories require special mechanisms for linking thematic roles to the sentence's syntax. It may be these linking mechanisms, housed in the frontal cortex, that are damaged.

Thus, the scope of the relations represented in lmSTC is unclear. The present results, in conjunction with the previous literature, demonstrate that lmSTC represents the values of semantic role variables for two-place verbal predicates describing events. Whether it is more generally involved in composing structured representations involving relations (e.g., spatial, stative) and composing the meaning of one-place predicates (e.g., "walked") are important topics for future research. There is some evidence to suggest that lmSTC encodes relational states as well events (Wu, Waller & Chatterjee, 2008; Dronkers et al., 2004), and one one-place as well as two-place predicates (Dronkers et al., 2004). But this evidence is weaker than that implicating it in the representation of two-place predicates denoting events.

A number of recent studies by Pylkkanen and colleagues have implicated a region of the left anterior temporal lobe in simple adjective-noun semantic composition; for example, integrating the words ‘red’ and ‘boat’ to represent a red boat (Westerlund et al., 2015; Pylkkanen et al., 2014; Bemis & Pylkkanen, 2011). Linguists consider adjective/noun composition to be a different type of semantic composition than the composition required to encode who did what to whom (Heim & Kratzer, 1998). The composition-type studied by Pylkkanen et al., is known as “predicate modification”, while the latter, following Frege, requires “function application”. These composition-types differ in the rules by which the truth conditions of an expression are derived. Nonetheless, it is intriguing that both may converge on areas of the left mid-anterior temporal lobe. Pylkannen et al. uses MEG, which has considerably lower spatial resolution than fMRI, making it hard to locate their region in relation to the region of present interest. It appears to be somewhat anterior, however. This would be consistent with a recent study by Baron & Osherson (2012), who found a region of left anterior superior temporal cortex to carry information about adjective/noun conceptual combinations (But see Price et al., 2015 for recent evidence implicating the left parietal lobe in predicate modification). Although predicate-modification and function-application may involve the application of different rules, it would not be surprising if they were performed in nearby parts of cortex. Both should take as input lexico-semantic representations and syntactic information, and produce, as output, complex, semantic structures. Regions that perform these operations need to receive input from many of the same cortical regions, and their outputs should, one would think, be read by the same cortical regions. Putting computations that have similar input and outputs constraints together would be an energy efficient solution for the brain.

Although we believe these studies provide compelling evidence that lmSTC represents semantic role variables, we do not assume that lmSTC is the only region involved in representing the meaning of sentences involving an agent, event-type, and patient. Of the three studies presented, Experiment 1 is the only to use a whole-brain analysis, and hence provided the only opportunity to identify regions outside lmSTC. Indeed, the dorsal-most part of this ROI entered the inferior-most part of the parietal lobe. However, we found no specialization for agent, patient, or verb information in the parietal portion of this region. We targeted regions with a very specific representational profile, searching for regions in which the patterns of activity (1) carry information about the different values of semantic variables, (2) allow active and passive versions of variable/value combinations to be grouped together and (3) are invariant across classes of semantically similar verbs. Any complex semantic representation that is contextually dependent upon specific verbs, specific variable-value interactions, or specific syntactic surface forms would not show up in this analysis.

Given that the present results were generated using only visually presented sentences, the current data are silent as to whether these representations are part of a general language of thought, or whether they are specifically linguistic. In particular, we don’t know whether we would obtain similar results using alternative modes of presentation, such as pictures. We note that Wu, Waller, & Chatterjee (2008) report deficits in comprehension of pictorial stimuli following damage to this region. However, linguistic deficits could disrupt comprehension of pictures if pictorial information is normally translated into words. Finding a phonological coding scheme in Experiment 3 would have provided evidence that this system is specifically linguistic. We find no such evidence, however, despite reports that nearby regions respond to phonological contrasts (Vigneau et al., 2006). The patterns of activity could still represent words, however: just using a non-phonological representational format.

We view the failure to find phonological or semantic similarity structure within these regions as an intriguing null effect. We take seriously the possibility that lmSTC does, in fact, use a code that might appear arbitrary to the observer. Such a code could have a number of advantages. Random or pseudorandom coding schemes ensure that symbols are easily discriminable from one another (Gallistel & King, 2011), which here, would make it easy for other neural systems to decode who did what to whom. In the introduction, we detailed how Smolensky's (1990) tensor product model needs either the vectors encoding the variables, or the vectors encoding the values to be statistically independent. Otherwise, there is no procedure through which an outside system can faithfully recover the identity of the constituent representations in a tensor product[11]. Arbitrary codes are thus useful in ensuring that the system exhibits compositionality, at least in the tensor product model.

Apparent randomness in the code is also a mark of efficient coding (Hopfield, 1982). In a conventional computer, data compression eliminates statistical redundancy,

[11] However, we find evidence that the representation of the variables is linearly independent in virtue of being represented using different populations of neurons (e.g., [111 000], for agent and [000 111] for patient. So, strictly, it would not be necessary for the representation of the set of values to be linearly independent as well.

which entails eliminating the overlap between data structures. Data that has been compressed for efficient storage can thus appear random to an outside observer, even if the encoding owes to a deterministic process. Given the number of bindings that need to be created, and the issue of scalability that hovers over models of binding, it would not be surprising if natural selection has favored efficient, compressed codes for encoding variable/value combinations. Similarity-based codes redundantly represent non-distinguishing information across data structures to facilitate generalization between representations. This is precisely the information that a compression algorithm could remove for efficiency. A compressed code is therefore at odds with the similarity-based coding ubiquitous in connectionist models.

If the representations in lmSTC are pointers to locations elsewhere in cortex (Kriete et al., 2013), similarity-based generalization could be performed in the cortical regions housing the symbols to which lmSTC points. The patterns of activity in lmSTC could be succinctly written addresses to these locations, which house richer representations. If so, we may be able to use pattern-based effective connective analysis to determine what brain regions lmSTC reads from, and how particular patterns in lmSTC affect/are affected by the patterns elsewhere representing this symbolic content. Posterior left middle temporal gyrus is a region of particular interest. This area, broadly defined, is involved in lexical access and word-level comprehension (Hickok & Poeppel, 2007; Dronkers et al., 2004), represents the semantic similarity of both words and pictures (Fairhall & Camarazza, 2012), and may represent information about a verb's arguments (Peelen et al., 2012). It's thus possible that the basic symbolic content that needs to be combined and recombined in lmSTC is housed in MTG. A functional relationship between these two regions is consistent with known anatomical and functional connections. Turken & Dronkers (2011) report functional coupling between the mid-anterior superior temporal gyrus/sulcus and the posterior middle temporal gyrus at rest (Turken & Dronkers, 2011), likely through the middle longitudinal fasciculus. If lmSTC represents pointers, posterior middle temporal gyrus is a strong candidate region for where it points.

It is important to bear in mind, however, that even if the encoding is apparently arbitrary (which we do not have strong evidence for at the moment), our analyses demonstrate that the symbols are stable across tokenings. That is, even if the mapping from (e.g.,) a noun meaning to a pattern of activity is random across the entire domain, the mapping from a noun meaning to a pattern of activity *for a given symbol, across tokens* is stable. Functionally, this stability is a necessary precondition for the symbol having a determinate meaning within the operation of the system, and for complex symbolic expressions to be composed using the same parts. Statistically, this stability is a necessary precondition for the classifiers to succeed in Experiments 2 and 3.

However, at this point, this is just an intriguing null result for the particular similarity models we test. To truly understand the code, it will be important to test a much wider range of stimuli with a much wider range of models, and with more statistical power. For, in evaluating the significance of the present results, we note that the classification accuracies observed here are rather modest, which weakens the null effect on the similarity models. In general, we regard the effects observed in these three studies as significant, not because of their size, but because they provide evidence for a distinctive theory of how the brain represents sentence meaning. In multiple studies, we find evidence for a functional, and corresponding anatomical, segregation based on semantic role, which may enable the composition of complex semantic representations. Such functional segregation need not take the form of spatial segregation, but insofar as it does involve spatial segregation, it becomes possible to provide evidence for functional segregation using fMRI, as done here.

The binding problem and lmSTC

In addition to providing evidence that the brain represents the values of abstract semantic role variables, our results bear on how the brain solves this particular binding problem. Given the temporal limitations of fMRI, the current design cannot provide direct evidence for or against temporal synchrony models. Our data suggest that temporal correlations are unnecessary, however. If semantic variables are represented by distinct neural populations, and the values of these variables are signaled by patterns of activity within each region, there is no need for a special mechanism to synchronize or integrate this information from disparate populations[12]. The brain maps these reusable variables, and in doing so, avoids a neural binding problem at this level, contrary to common assumption (Hummel et al., 2000; Doumas et al., 2008; Stewart & Eliasmith, 2008; Van der Velde, 2006).

This representational strategy is not unique to semantic role variables. Analogous solutions can be seen in other representational problems that the brain faces. For example, early visual cortex has an analogous solution to representing the visual field. In striate cortex, portions of the visual field (the variables[13]) are represented by separate neural populations. Cells that are sensitive to particular line orientations (the values), are

[12] That is, there is no special problem of integration *at this level of representation.* Information about the agent and patient must then be integrated to reason about the entire event. Our analyses in Experiments 2 and 3 targeted representations that are reused across propositions, rather than representations that are unique to particular proposition-level content.

[13] The application of the notion 'variable' is more apt for mid-high level cognition, where there is reason to believe mental algorithms explicitly reference variables in their operation. I don't know whether there are algorithms that reference 'upper-left hemifield' as such, but I believe the comparison is still instructive.

organized in columns and hypercolumns, and, critically, duplicated for each portion of each visual field. Each orientation column represents the presence of a line orientation at that particular position in the visual field, thus representing a conjunction of features. The variables are mapped across the cortex, individuated by their anatomical location. The feature-values (e.g., a particular line orientation) are duplicated for each variable-specific sub-region. The content at different parts of the visual field can vary independently of the content at another, and this architecture allows different variable/value pairs to be simultaneously and independently represented.

Our data suggests that, at a high-level, it may be likewise for variable/value binding in lmSTC. Many events have a common semantic structure, such that one can identify abstract recurring variables, such as the agent or patient, whose values vary independently of one another. These semantic variables are mapped across lmSTC, again individuated by their anatomical location. Representations of the values (the specific content of which is still unclear) appear to be duplicated for the different variables, just as line-orientations are duplicated. Cast in this light, mapping can be seen as a general strategy for flexibly representing reusable variables that can take different values, separately and simultaneously[14]. Other cortical systems that need to integrate information about the identity of the agent and patient can read the values of these variables at the same time. Of course, given that we see no similarity structure in lmSTC, there may be a dis-analogy in the actual coding schemes used to represent variable/value combinations. In early visual cortex, adjacent columns code for similar line orientations,

[14] We do not assume, however, that every neural map arises from these particular demands; only that given this set of demands, mapping and duplication emerge as a plausible solution. For example, the recently discovered map of numerosity in the parietal lobe (Harvey et al., 2013) likely serves different purposes.

meaning similar orientations are encoded by similar patterns of activity. Thus far, we have seen no similarity structure in lmSTC.

Mapping variables is thus a general and elegant solution for the brain to take across domains. But is it wasteful? Why would the brain redundantly encode the same symbol in every role in which it can occur? It is one thing for the brain to duplicate representations of line orientations, of which there is a relatively small number, but something quite different to duplicate all the nouns or corresponding concepts one knows (or the addresses thereof). Some have doubted that the brain would be organized in such an apparently wasteful way (Hinton, 1986; Haysworth, 2012; Kriete et al., 2013). Although it's impossible to determine the feasibility of the solution without knowing what the patterns in lmSTC represent, a few simple estimates suggest that any cortical region the size of lmSTC is likely up to the challenge.

Braitenberg & Schuz (1991) report that mouse cortex contain ~92,000 neurons per mm^3 in a generic region of cortex. Koch (unpublished) uses this infer that primate brains have around ~100,000 neurons/ mm^3. We have no reason to believe lmSTC is any different, so we assume this holds here as well. Although we do not know what, exactly, lmSTC represents, we will assume as a first pass it is words, or at least not so different from the number of words we know. Pinker (1994) estimates that the average adult knows around 60,000 words. Of these, Hudson (1994) takes the percentage of nouns in various large corpora of English to be ~37%. Given an inefficient code in which the firing of one and only one neuron represents one and only one filler, there is still, in principle, enough information in the space sampled by one 1.5 mm^3 voxel to encode all the nouns that the average adult knows. In fact, using these estimates, there would be enough neurons in a single voxel to represent all the nouns 10 adults know. The group-level ROIs we identified in Experiment 2 have not 1, but 60 and 180 voxels for the agent and patient regions, respectively. And a distributed code increases the representational capacity of the region exponentially. At a maximum, the 100,000 neurons in 1 mm^3, treated as binary units, could encode $2^{100,000}$ symbols. This, of course, is just as implausible as the one symbol/one neuron code; as we reviewed in the introduction, such coarse codes are wasteful, given the high cost of signaling relative to the metabolic cost of maintaining a neuron (Laughlin et al, 2001). This is just to say, however, that some duplication and reduplication of the representations of various symbols (or perhaps even more likely, addresses for those symbols) is not as physically problematic as other authors have suggested (Hinton, 1986; Haysworth, 2012; Kriete et al., 2013).

But the plausibility of this arrangement depends not just on the number of fillers, but also on how many times they would need to be duplicated. How many semantic role variables are there? This is particularly difficult to answer. Early work on semantic roles in linguistics treated semantic roles as lists of unanalyzable primitives, with no internal or external structure (see Levin & Rappaport-Hovav, 2005 for review). For example, the online verb-lexicon VerbNet (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html) uses 23 distinct semantic role labels, from the familiar agent and patient, to highly specific roles like "asset" used only in alternations of a sum of money. Van Valin (2005) lists 39 such roles. If we were to repeat our experiments with the 23 different semantic roles on VerbNet, it seems unlikely that we would find a distinct sub-region in lmSTC for 'asset'. Unfortunately, when semantic roles are unstructured lists, there appears to be no principled way of telling just how many roles there are, or, more importantly for present purposes, why roles covary; for example, why verbs that have a patient often have an agent, but not an experiencer (Levin & Rappaport-Hovav, 2005). An unstructured list of semantic roles is also insufficient to explain the empirical facts of our final analysis in which we see patterns of activity that generalize across stimulus and agent, and patient and experiencer, but not other combinations.

At the other extreme, the barest conceptions of semantic roles posit two generalized role-buckets (e.g. "actor/undergoer"), into which narrower roles can be sorted (Van Valin, 1997; 2005; Dowty, 1991). Although these representations might exist elsewhere, this is not what lmSTC represents. We saw in Experiment 3 that there are no sub-regions that group the various actors together, and there are topographic differences between agent and patient, and stimulus and experiencer that need to be explained. So despite that fact that it would be highly feasible to represent only two variables given the physical constraints of lmSTC, this is not consistent with the evidence.

One appealing alternative is to decompose verb-meanings into simpler, recurring semantic elements (Jackendoff, 1972; 1992; Pinker, 1989; Rappaport & Levin, 1988). These primitive elements could themselves be argument-taking functions, such as (e.g., ([x ACT]) and can be combined to form more complex structures (e.g., to represent a causal event [[x ACT] CAUSE [BECOME [y *STATE*]]. On this approach, semantic roles are actually labels for argument positions in a verb's semantic decomposition. These basic functions then recur across verbs. We mentioned this alternative in the discussion of Experiment 3, citing Jackendoff's 'GO(thing, place)' function as one that could potentially explain our results, if it is interpreted as referring to the conceptual structure of the event more broadly. The particular constellation of semantic features that define verb-meanings then further depend upon which events can actually occur, and which can not (Levin & Rappaport-Hovav, 2005). Grounding a theory of verb meaning in a theory of possible events provides a principled basis for the set of semantic variables, though the best conceptualization of event structure is a matter of debate (Levin & Rappaport-Hovav, 2005).

Levin & Rappaport-Hovav (2005) note that that number of semantic primitives in predicate-decomposition models of verb meaning is typically smaller than the number of semantic roles in role-lists. However, they appear to be roughly the same order of magnitude. In one case, Van Valin (2005; pg 55, Table 2.4) reduces 39 distinct semantic roles to ~30 primitive argument positions. A preference for predicate-decomposition models thus owes more on their explanatory adequacy in linguistics than to their ability to reduce the number of roles a theory needs to posit. In addition to providing a better understanding of the regularities at the syntax-semantics interface, we see in Experiment 3 that predicate-decompositions are also, perhaps not coincidentally, more likely to explain the pattern of results in lmSTC. Theories that specify basic, semantic building blocks with function-argument structure (e.g., ACT, BE, GO, STAY, CAUSE) have the resources to explain why we see certain semantic roles (e.g., agent/stimulus), but others not (patient/stimulus) group together[15]. Although, at present, we should be cautious about using lmSTC to test linguistic theories, we can fruitfully do the converse: use different models of semantically primitive functions to test what semantic groupings lmSTC respects. This is an exciting direction for future work.

[15] That is, they have the resources in principle, though we cannot say which are driving the results of Experiment 3. This is a limitation of the stimuli used, and not the theoretical framework, however. See Discussion in Experiment 3.

But, to return to the issue of physical plausibility, how many variables *could* lmSTC represent? Would 40 be too many, should there be that many semantically primitive functions? What about 100, or 1000? The region we searched in Experiment 3 has ~600 voxels, each with a volume of 8mm^3. This region of the left superior temporal gyrus and sulcus thus has a volume of approximately 4,800mm^3. Based on Koch's estimate, this region has ~480,000,000 neurons with which to represent different variable/value combinations. Therefore, if every neuron represented one and only one noun, this region could still represent ~12,000 variables. (((100,000 neurons/mm^3*4,800mm^3)/40,000nouns)=12,000 variables).

These particular quantities should not be taken seriously. We do not know what, exactly lmSTC represents, or how neurons represent it. It's possible the representations can be more complex than a single noun (e.g., 'the small elephant with the blue eye patch chased the dog'). This is only to emphasize that representing 10, 50, or even 500 variables, each one duplicating representations for all the possible filler values is likely physically possible, even if we assume an inefficient coding scheme. This is not intuitively obvious, and has led writers to doubt such a solution for variable/value binding (Hinton, 1986; Haysworth, 2012; Kriete et al., 2013). We should not rule out this solution due to concerns about plausibility.

However, there are still a number of outstanding theoretical issues outlined in the introduction that the present data do not speak to**.** The present studies provide no evidence regarding what procedures are used to compute these bindings. This is not a principled problem, however. This computation may be performed through rule-like matrix-operations, such as the tensor product (Smolensky, 1991) or circular convolution models

(Plate, 1994). Or they could be computed using a more complicated serial architecture with short-term memory (Kriete et al., 2013). Our data suggest that once they are computed, however, the outputs are represented as patterns of activity over variable-specific neural populations. Such a map provides a physical basis through which the constituent structure of a complex expression can be encoded. Second, the regions we discover are likely to be at a very early stage of combinatorial semantic representation. Other cortical systems need to integrate information from these variable-specific sub-regions in order to reason about an event in its entirety. This architecture does not solve the problem of how the brain encodes singular events. Thus, although the present data address one level at which there is an alleged binding problem, they do not address the binding problem in its entirety. Third, our data provide no evidence regarding how multiple bindings using the same predicates are encoded simultaneously. We do not show how, to return to our earlier example, the brain encodes "John loves Mary, but Mary loves Jesus". Our data suggest that each variable is individuated by its anatomical location. But the fact that we can understand who loves whom in the above propositions suggests that there is not simply (e.g.,) one agent register and one patient register. It's possible that lmSTC has multiple registers per variable. There need not be a one-to-one mapping. Based on the above estimate, the anatomical location of each variable could be sufficiently large to allow for multiple variable-specific registers to be arranged within it. Of course, it would still then remain to be seen how the representations in these separate registers would be differentially conjoined within propositions. If these regions contain sets of registers, it is possible that particular registers in, for example, the agent region are directly linked to particular registers in the patient region, mirroring the structure we see at a high-level. This is an important question for future research, and one the present data do not speak to.

Finally, our data provide no evidence for how bindings are stored in long-term memory (Marcus, 2001; Hummel & Holyoak, 1997; Doumas et al., 2008). While it is possible, in principle, for registers to store data in long-term memory (Marcus, 2001), it's less clear how this should work in our current model. Some models rely on two-systems: one for binding in short-term memory and one for long-term storage of conjunctions. For example, Hummel et al.'s model (Hummel & Holyoak, 1997; Doumas et al., 2008) uses temporal synchrony to transiently encode bindings in short-term memory, and then a conjunctive coding scheme to store them in long-term memory. Although this is not the most parsimonious explanation, such a two-system approach may ultimately be the right one (although we disagree with the specifics of the binding-construction system). The field of cognitive neuroscience generally assumes that the medial temporal lobes must bind disparate information into conjunctive representations to support long-term memory: for example, of particular sights, smells, and sounds. To a first approximation, the problem appears no different in storing details about who did what to whom than it does in storing these other conjunctive details. It merely conjoins a different class of inputs. Given that our working model of the brain suggests that the medial temporal lobe stores conjunctive representations from other inputs, it imposes no additional (principled) burden to assume it does so for representations of who did what whom. How the medial temporal lobes store this information in physical form is, of course, a mystery (See Gallistel & King, 2011). We suggest that the lmSTC is critically involved in constructing the initial representations of who did what to whom in an event, however higher-order representations of the event are ultimately stored.

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

References

Ansell, B. J., & Flowers, C. R. (1982). Aphasic adults' use of heuristic and structural linguistic cues for sentence analysis. *Brain and Language*, *16*(1), 61-72.

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, *21*(10), 1133-1145.

Baker, M. C. (1997). Thematic roles and syntactic structure. In *Elements of grammar* (pp. 73-137). Springer Netherlands.

Baron R.M., & Kenny D.A., (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51(6):1173-1182.

Baron S.G., Thompson-Schill S.L., Weber M., & Osherson D., (2010) An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cognitive Neuroscience* 1(1):44-51.

Baron S.G., & Osherson D. (2011) Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage* 55(4):1847-1852.

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, *6*(3), 291-352.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17-26.

Bemis D.K., & Pylkkanen L. (2011) Simple Composition: A Magnetoencephalography Investigation into the Comprehension of Minimal Linguistic Phrases. *Journal of Neuroscience* 31(8):2801-2814.

Berndt, R. S., Mitchum, C. C., & Haendiges, A. N. (1996). Comprehension of reversible sentences in “agrammatism”: A meta-analysis. *Cognition*, *58*(3), 289-308.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767-2796.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annu. Rev. Psychol.*, *58*, 47-73.

Braitenberg, V., & Schüz, A. (1991). *Anatomy of the cortex: Statistics and geometry*. Springer Verlag Publishing.

Caramazza, A., & Miceli, G. (1991). Selective impairment of thematic role assignment in sentence processing. *Brain and Language*, *41*(3), 402-436.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, *13*(11), 1428-1432.

Chiu Y.C., Esterman M.S., Gmeindl L., & Yantis S., (2012) Tracking cognitive fluctuations with multivoxel pattern time course (MVPTC) analysis. *Neuropsychologia* 50(4):479-486.

Churchland, P. M. (1996). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. MIT Press.

Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y. C., & Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience*, *32*(8), 2608-2618.

Coutanche M.N., & Thompson-Schill S.L., (2013) Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in human neuroscience* 7.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and language*, *3*(4), 572-582.

Cox R.W., (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29(3):162-173.

Dapretto, M., & Bookheimer, S. Y. (1999). Form and content: dissociating syntax and semantics in sentence comprehension. *Neuron*, *24*(2), 427-432.

Devauchelle A.D., Oppenheim C., Rizzi L., Dehaene S., & Pallier C. (2009) Sentence syntax and content in the human temporal lobe: an fMRI adaptation study in auditory and visual modalities. *Journal of cognitive neuroscience* 21(5):1000-1012.

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

Doumas L.A., Hummel J.E., & Sandhofer C.M., (2008) A theory of the discovery and predication of relational concepts. *Psychological review* 115(1):1-43.

Dowty, D. (1991). Thematic proto-roles and argument selection. *language*, 547-619.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1), 145-177.

Duda R., Hart P.E., & Stork D.G. (2001) *Pattern classification* (Wiley, New York) 2nd Ed.

Efron B. & Tibshirani R. (1993) *An introduction to the bootstrap* (Chapman & Hall, New York) pp xvi, 436 p.

Eger E., Ashburner J., Haynes J.D., Dolan R.J., & Rees G. (2008) fMRI activity patterns in human LOC carry information about object exemplars within category. *Journal of cognitive neuroscience* 20(2):356-370.

Eisenberger N.I., Lieberman M.D., & Williams K.D. (2003) Does rejection hurt? An FMRI study of social exclusion. *Science* 302(5643):290-292.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *The Journal of Neuroscience*, *33*(25), 10552-10558.

Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177-1194.

Fedorenko E, Behr M.K., & Kanwisher N. (2011) Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America* 108(39):16428-16433.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Clarendon Press.

Fodor J.A., & Pylyshyn Z.W., (1988) Connectionism and Cognitive Architecture - a Critical Analysis. *Cognition* 28(1-2):3-71.

Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183-204.

Fodor, J. (1997). Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition*, *62*(1), 109-119.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). " Who" Is Saying" What"? Brain-Based Decoding of Human Voice and Speech. *Science*, *322*(5903), 970-973.

Frege, G. (1892). On Concept and Object. *Translations from the Philosophical Writings of Gottlob Frege*, Geach and Black (eds.), Basil Blackwell, Oxford, 1977, pp. 42–55.

Friederici A.D., Ruschemeyer S.A., Hahne A., & Fiebach C.J., (2003) The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral cortex* 13(2):170-177.

Friston K.J., (2011) Functional and effective connectivity: a review. *Brain connectivity* 1(1):13-36.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, *7*(2), 77-83.

Gallistel, C. R., & King, A. P. (2011). *Memory and the computational brain: Why cognitive science will transform neuroscience* (Vol. 6). John Wiley & Sons.

Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current opinion in neurobiology*, *16*(2), 240-246.

Haxby J.V.*, et al.* (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539):2425-2430.

Hagoort P., Hald L., Bastiaansen M., & Petersson K.M., (2004) Integration of word meaning and world knowledge in language comprehension. *Science* 304(5669):438-441.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, *9*(9), 416-423.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual review of neuroscience*, *37*, 347-362.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194-205.

Han, Z., Bi, Y., Chen, J., Chen, Q., He, Y., & Caramazza, A. (2013). Distinct Regions of Right Temporal Cortex Are Associated with Biological and Human–Agent Motion: Functional Magnetic Resonance Imaging and Neuropsychological Evidence. *The Journal of Neuroscience*, *33*(39), 15442-15453.

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. *Science*, *341*(6150), 1123-1126.

Hayworth, K. J. (2012). Dynamically partitionable autoassociative networks as a solution to the neural binding problem. *Frontiers in computational neuroscience,6*.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar* (Vol. 13). Oxford: Blackwell.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393-402.

Hinton, G. E. (1986, August). Learning distributed representations of concepts. *Proceedings of the eighth annual conference of the cognitive science society*, 1, 12.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language*, 331-339.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, *110*(2), 220.

Hummel, J. E., Holyoak, K. J., Green, C., Doumas, L. A., Devnich, D., Kittur, A., & Kalar, D. J. (2004). A solution to the binding problem for compositional connectionism. In *Compositional connectionism in cognitive science: Papers from the AAAI Fall Symposium, ed. SD Levy & R. Gayler* (pp. 31-34).

Humphries C, Binder JR, Medler DA, & Liebenthal E (2006) Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of cognitive neuroscience* 18(4):665-679.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar* (pp. 76-ff). Cambridge, MA: MIT press.

Jackendoff, R. (1990). *Semantic structures* (Vol. 18). MIT press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Kamitani Y., & Tong F., (2005) Decoding the visual and subjective contents of the human brain. *Nature neuroscience* 8(5):679-685.

Kay K.N., Naselaris T., Prenger R.J., & Gallant J.L., (2008) Identifying natural images from human brain activity. *Nature* 452(7185):352-355.

Koch, C. (unpublished manuscript). Neuronal and Syntaptic Packing Densities. Retrieved from centrosome.caltech.edu/courses/cns187/references/neuronal_densities.pdf, Feb. 2015.

Kriegeskorte, N., Goebel R., & Bandettini, P. (2006) Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America* 103(10):3863-3868.

Kriegeskorte N., Formisano E., Sorger B., & Goebel R. (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 104(51):20600-20605.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126-1141.

Kriete T., Noelle D.C., Cohen J.D., & O'Reilly R.C. (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences of the United States of America* 110(41):16390-16395..

Kuperberg, G., McGuire, P., Bullmore, E., Brammer, M., Rabe-Hesketh, S., Wright, I., & David, A. (2000). Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. *Cognitive Neuroscience, Journal of*, *12*(2), 321-341.

Kuperberg, G., Holcomb, P., Sitnikova, T., Greve, D., Dale, A., & Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Cognitive Neuroscience, Journal of*, *15*(2), 272-293.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205.

Laughlin, S. B., van Steveninck, R. R. D. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature neuroscience*, *1*(1), 36-41.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, *11*(4), 475-480.

Linebarger, M. C., Schwartz, M. F., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition*, *13*(3), 361-392.

Ledoit O. ,& Wolf M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88(2):365-411.

Ledoit O. & Wolf M. (2002) Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics* 30(4):1081-1102.

Lennie, P. (2003). The cost of cortical computation. *Current biology*, *13*(6), 493-497.

Levin, B., & Hovav, M. R. (2005). *Argument realization*. Cambridge University Press.

Love, B. C. (1999). Utilizing time: Asynchronous binding. *Advances in Neural Information Processing Systems*, 38-44.

MacKinnon D.,P., Lockwood C.M.,, & Williams J., (2004) Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research* 39(1):99-128.

Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: A cognitive neuropsychological perspective. *Annual review of psychology*, *60*, 27.

Mahon B.Z, Anzellotti S., Schwarzbach J., Zampini M., & Caramazza A (2009) Category-specific organization in the human brain does not require visual experience. *Neuron* 63(3):397-405.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive psychology*, *37*(3), 243-282.

Marcus G.F., (2001) *The algebraic mind: Integrating connectionism and cognitive science* (The MIT Press).

Martin, K. A. (1994). A brief history of the "feature detector". *Cerebral cortex*,*4*(1), 1-7.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., & Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467-479.

Meltzer, J. A., McArdle, J. J., Schafer, R. J., & Braun, A. R. (2010). Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. *Cerebral cortex*, *20*(8), 1853-1864.

Miozzo, M., Rawlins, K., & Rapp, B. (2014). How verbs and non-verbal categories navigate the syntax/semantics interface: Insights from cognitive neuropsychology. *Cognition*, *133*(3), 621-640.

Mitchell T.M.*, et al.* (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191-1195.

Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1353.

Norman K.,A, Polyn S.M, Detre G.J, & Haxby J.V. (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences* 10(9):424-430.

O'Reilly R.C., & Busby R.S., (2002) Generalizable relational binding from coarse coded distributed representations. *Advances in neural information processingsystems* 1:75 82.

Pallier C., Devauchelle A.D, & Dehaene S., (2011) Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the United States of America* 108(6):2522-2527.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976-987.

Peelen, M. V., Romagno, D., & Caramazza, A. (2012). Independent representations of verbs and actions in left lateral temporal cortex. *Journal of cognitive neuroscience*, *24*(10), 2096-2107.

Pesetsky, D. (1987). Binding problems with experiencer verbs. *Linguistic Inquiry*, 126-140.

Phelps E.A., & LeDoux J.E., (2005) Contributions of the amygdala to emotion processing: f rom animal models to human behavior. *Neuron* 48(2):175-187.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 193.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT press

Pinker, S. (1994). *The language instinct: The new science of language and mind* (Vol. 7529). Penguin UK.

Pinker, S. (1997). How the mind works. 1997. *NY: Norton*.

Plate T.A., (1995) Holographic reduced representations. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 6(3):623-641.

Poeppel, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D., & Braun, A. R. (2004). Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. *Neuropsychologia*, *42*(2), 183-200.

Polyn S.M., Natu V.S., Cohen J.D., & Norman K.A., (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310(5756):1963-1966.

Price, A.R., Bonner, M.F., Peelle, J.E., & Grossman, M. (2015). Converging Evidence for the Neuroanatomic Basis of Combinatorial Semantics in the Angular

Gyrus. *The Journal of Neuroscience*, *35*(7), 3276-3284.

Price, C. J., Moore, C. J., Humphreys, G. W., & Wise, R. J. S. (1997). Segregating semantic from phonological processes during reading. *Journal of Cognitive Neuroscience*, *9*(6), 727-733.

Pylkkänen, L., Bemis, D. K., & Elorrieta, E. B. (2014). Building phrases in language production: An MEG study of simple composition. *Cognition*, *133*(2), 371-384.

Newell, A. (1980). Physical Symbol Systems. *Cognitive science*, *4*(2), 135-183.

Rappaport, M., & Levin, B. (1988). What to do with Theta-Roles in Thematic Relations. *Syntax and semantics*, *21*, 7-36.

Richardson, F. M., Thomas, M. S., & Price, C. J. (2010). Neuronal activation for semantically reversible sentences. *Journal of cognitive neuroscience*, *22*(6), 1283-1298.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019-1025.

Rodd, J. M., Vitello, S., Woollams, A. M., & Adank, P. (2015). Localising semantic and syntactic processing in spoken and written language comprehension: An Activation Likelihood Estimation meta-analysis. *Brain and language*, *141*, 89-102

Rogalsky C., & Hickok G., (2009) Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral cortex* 19(4):786-796.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rothman, D. L., Sibson, N. R., Hyder, F., Shen, J., Behar, K. L., & Shulman, R. G. (1999). In vivo nuclear magnetic resonance spectroscopy studies of the relationship between the glutamate--glutamine neurotransmitter cycle and functional neuroenergetics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *354*(1387), 1165-1177.

Rozwadowska, B. (1989). Are thematic relations discrete?. *Linguistic categorization*, 115-130.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1988). *Parallel distributed processing* (Vol. 1, pp. 354-362). IEEE.

Schafer J., & Strimmer K., (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology* 4.

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon.

Schwartz, M. F., Saffran, E. M., & Marin, O. S. (1980). The word order problem in agrammatism: I. Comprehension. *Brain and language*, *10*(2), 249-262.

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, *24*(1), 67-77.

Shastri L., & Ajjanagadde V., (1993) From Simple Associations to Systematic Reasoning - a Connectionist Representation of Rules, Variables and Dynamic Bindings Using Temporal Synchrony. *Behavioral and Brain Sciences* 16(3):417-451.

Skeide M.A., Brauer J., & Friederici A.D., (2014) Syntax gradually segregates from semantics in the developing brain. *NeuroImage* 100:106-111

Sibson, N. R., Dhankhar, A., Mason, G. F., Rothman, D. L., Behar, K. L., & Shulman, R. G. (1998). Stoichiometric coupling of brain glucose metabolism and glutamatergic neuronal activity. *Proceedings of the National Academy of Sciences*, *95*(1), 316-321.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, *18*(1), 555-586.

Smolensky P. (1990) Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems. *Artificial Intelligence* 46(1-2):159-216.

Smolensky, P. (1991). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. In *Connectionism and the philosophy of mind*(pp. 281-308). Springer Netherlands.

Stephan K.E., & Friston K.J., (2010) Analyzing effective connectivity with fMRI. *Wiley interdisciplinary reviews. Cognitive science* 1(3):446-459.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2011). Neural representations of compositional structures: Representing and manipulating vector spaces with spiking neurons. *Connection Science*, *23*(2), 145-153.

Stewart, T.C., & Eliassmith, C. (2012). Compositionality and biologically plausible models. In Werning, M., Hinzen, W., & Machery, E. (Eds.). (2012). *The Oxford handbook of compositionality*. Oxford University Press.

Thothathiri, M., Kimberg, D. Y., & Schwartz, M. F. (2012). The neural basis of reversible sentence comprehension: Evidence from voxel-based lesion symptom mapping in aphasia. *Journal of cognitive neuroscience*, *24*(1), 212-222.

Todd M.T., Nystrom L.E., & Cohen J.D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage* 77:157-165.

Turken, U., & Dronkers, N. F. (2011). The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Frontiers in systems neuroscience*, *5*.

Vaden, K. I., Muftuler, L. T., & Hickok, G. (2010). Phonological repetition-suppression in bilateral superior temporal sulci. *Neuroimage*, *49*(1), 1018-1023.

Valiant, L. G. (2012). The hippocampus as a stable memory allocator for cortex. *Neural computation*, *24*(11), 2873-2899.

Van der Velde, F., & De Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(01), 37-70.

Van Valin, R. D. (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press.

Vandenberghe, R., Nobre, A., & Price, C. (2002). The response of left temporal cortex to sentences. *Cognitive Neuroscience, Journal of*, *14*(4), 550-560.

Vickery T.J., Chun M.M., & Lee D. (2011) Ubiquity and Specificity of Reinforcement Signals throughout the Human Brain. *Neuron* 72(1):166-177.

Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, *30*(4), 1414-1432.

Von Der Malsburg, C. (1994). *The correlation theory of brain function* (pp. 95-119). Springer New York.

Von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. *Neuron*, *24*(1), 95-104.

Wager T.D., Davidson M.L., Hughes B.L., Lindquist M.A., & Ochsner K.N. (2008) Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron* 59(6):1037-1050.

Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and language*, *141*, 124-134.

Wu D.H., Waller S., & Chatterjee A., (2007) The functional neuroanatomy of thematic role and locative relational knowledge. *Journal of cognitive neuroscience* 19(9):1542-1555.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences*, *104*(15).

Appendix: Detailed Experimental Methods, and Supporting Information.

Experiment 1 Supporting Information

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

Data Acquisition and Preprocessing. 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 (1.2 mm^3 isotropic voxel MPRAGE) was collected prior to functional data acquisition. The echo-planar imaging (EPI) pulse sequence for functional scans employed a 2500 ms TR, a TE of 30 ms, and flip angle of 85°. Each volume consisted of 39 3-mm slices parallel to the AC-PC plane. Stimuli were presented using Psychtoolbox software (http://www.psychtoolbox.org) for Matlab (http://www.mathworks.com).

Image preprocessing was performed using a combination of AFNI functions and custom scripts. 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.

The data used in pattern-based effective connectivity analyses (PBEC) underwent several further preprocessing operations. For both predictor and criterion data, the mean signal level at each TR in the lateral ventricles and a white matter ROI were regressed out of the entire volume. This was done to remove interregional dependencies attributable to non-cognitive, physiological coupling, as is common in functional connectivity analyses (36). Additionally, the criterion, but not predictor data were smoothed with a Guassian kernel at 6mm FWHM. This served to enhance the signal-to-noise ratio in those regions in which we were interested in mean signal level across the entire region, while preserving finer spatial patterns in predictor regions. Thus, we used unsmoothed data in lmSTC to predict smoothed data in the left amygdala ROI.

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

Experiment 1 – Supporting Materials, Methods, and Results

			Use in Exp 1.
1	**The grandfather kicked the baby**	**The baby kicked the grandfather**	**connectivity**
2	**The mother struck the boy**	**The boy struck the mother**	**connectivity**
3	The father pulled the child	The child pulled the father	ROI localization
4	The grandmother touched the girl	The girl touched the grandmother	ROI localization
5	The truck hit the ball	The ball hit the truck	ROI localization
6	The door smacked the branch	The branch smacked the door	ROI localization

			Use in Exp. 1
1	**The baby was kicked by the grandfather**	**The grandfather was kicked by the baby**	**connectivity**
2	**The boy was struck by the mother**	**The mother was struck by the boy**	**connectivity**
3	The child was pulled by the father	The father was pulled by the child	ROI localization
4	The girl touched the grandmother	The grandmother touched the girl	ROI localization
5	The ball was hit by the truck	The truck was hit by the ball	ROI localization
6	The branch was smacked by the door	The door was smacked by the branch	ROI localization

Table S1. Active and passive versions of the sentences employed in Experiment 1.

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

Whether the proposition was presented in the active or passive voice on a given trial was randomly determined. Active and passive versions of the same proposition were treated identically for all analyses. Three strings of non-words were also presented to subjects in each run, but were not analyzed. On one third of the trials, questions were presented following the fixation period. These consisted of questions about the agent of the immediately preceding proposition (e.g., "did the ball hit something?"), questions about the patient ("did the ball get hit by something?"), or prompts to rate "how morally bad" the event was on a scale of 1-5, with one being "not bad at all" and 5 being "very bad." 50% of the comprehension questions had affirmative answers. The subjects' responses were signaled using a right-hand button box. Which question was presented on a given trial was randomly determined, as were the particular trials that were followed by comprehension questions. These were included simply to promote subject engagement, and were not analyzed.

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

We conducted our searchlight analyses using the Searchmight Toolbox (33). A cube with a 2-voxel (6mm) radius was centered at each voxel, and a linear discriminant classifier with a shrinkage estimate (35) of the shared population covariance matrix was used to probe the surrounding region for informational content. Non-edge neighborhoods contained 124 voxels. Pair-wise classifiers were separately trained for each of the four mirror-image proposition pairs. For every pair, performance at a given location was assessed by iteratively holding out each run as test data, training the pair-wise classifier on 12/13 runs, testing on the held-out run, and averaging performance across the 13 cross-validation folds. This resulted in a single whole-brain accuracy map per mirror-image proposition pair. These four pair-level maps were then averaged to create one map across pairs for each subject, with the aim of identifying regions that consistently contained information across the four mirror-image pairs.

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

We first chose a voxelwise threshold of p<0.005 and a corrected threshold of p<0.05, but no statistically significant clusters survived this threshold. Given that our chief aim was to simply localize candidate regions for Experiment 2, we therefore lowered the voxelwise threshold to p<0.02. At this threshold, two ROIs were found to be highly statistically significant, even when correcting for testing multiple voxelwise thresholds (lmSTC, 123 voxels, clusterwise p<0.0001: right EC, 108 voxels, p<0.001 clusterwise). The principal drawback to using this liberal voxelwise threshold is that the statistical significance of these regions owes largely to the size of the ROIs. This makes it unlikely that the discovered regions constitute homogenous functional units. Experiment 2 addresses this concern by probing the informational content of one of these two regions.

The first discovered ROI was located in the left mid superior temporal cortex (lmSTC). It begins in the inferior-most part of the parietal cortex, extends through the mid portion of the superior temporal gyrus and superior temporal sulcus and terminates in the middle temporal gyrus (See Supplementary Figure 1A). The second ROI is centered on the right posterior insula/extreme capsule fiber bundle. It encompasses parts of the posterior insula, claustrum, putamen, and medial-superior temporal lobe, (p<0.001, 108 voxels) (See Supplementary Figure 1B).

Experiment 1 Whole-Brain Searchlight Results

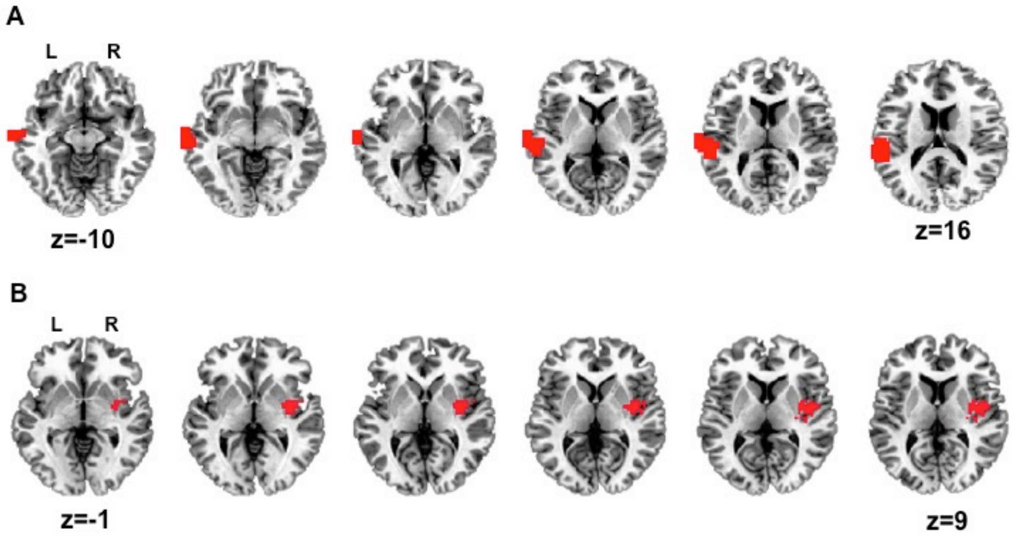

Figure S1 (continued). (A) left-mid superior temporal cortex (lmSTC) and (B) right insula/extreme capsule region discovered by Experiment 1 searchlight analysis distinguishing mirror-image propositions (e.g., “the truck hit the ball”/”the ball the truck”).

Post-hoc ROI analyses. As our principal searchlight results from Experiment 1 were obtained by averaging classification accuracy across four pairs, it is possible that the significance of the results owes to high accuracy on some subset of the pairs, with chance performance on the remaining pairs. To evaluate this possibility, we trained and tested linear discriminant classifiers with a shrinkage estimate of the covariance matrix separately on each pair in lmSTC to evaluate *post-hoc* which mirror-image proposition-pairs were driving our results. Supplementary Tables 2A and 2B show the results of these classifications by mirror-image proposition pair and ROI. A repeated-measures ANOVA revealed no significant differences in classification accuracy across the six pairs for either lmSTC ($F(5, 75)$=0.4, p=0.84) or the right insula/extreme capsule ROI (($F(5, 75)$=0.15, p=0.98). We find fairly consistent levels of classification accuracy in both ROIs, suggesting these regions are not driven by idiosyncracies of the particular pairs, or only by the animate or inanimate proposition-pairs. Instead, the pattern of results is consistent with both lmSTC and the right posterior insula/extreme capsule encoding domain-general information about “who did what to whom.”

If the regions discovered in the searchlight analysis do represent structure-dependent meaning, then they should facilitate classification of non-mirrored propositions as well. For example, they should be able to distinguish “the truck hit the ball” from “the father pulled the child.” Although these non-mirrored pairs are not well matched, and one would expect many other brain regions to be able to perform this classification (e.g., regions that encode the semantic/phonological content of the nouns and verbs), this analysis nevertheless serves as a “sanity check” on the ROIs localized using the searchlight analysis. To ensure that our ROIs could also discriminate non-mirrored pairs, pair-wise classifiers were separately used for the 24 non-mirrored comparisons that could be generated from the four mirror-image proposition pairs. This analysis was performed using data from the lmSTC ROI and the right posterior insula/extreme capsule ROI separately. These 24 classification accuracy statistics were then averaged for each ROI and submitted to a one tailed *t*-test against 0.5. Of the two ROIs able to discriminate within mirror-image proposition pairs, only the lmSTC ROI was able to reliably discriminate non-reversed pairings as well, t(15) = 4.06, p=0.005. The right extreme-capsule/insula ROI trended in this direction, but its results were not statistically significant, t(15) = 1.45, p<0.09. This failure to robustly classify non-mirror-image proposition pairs casts doubt on the possibility that the right posterior insula encodes complex, structured semantic representations. Taken in conjunction with other null results pertaining to this ROI, described in the “Pattern-Based Effective Connectivity” section below, we chose to focus on lmSTC in Experiments 2 and 3.

Right Insula	**Mirror-image pair**	**mean Accuracy**	***t*, one-tailed *p***
1	STRUCK (mother: boy)	.562	*t*=2.17, p=0.023
2	KICKED (grandfather: baby)	.557	*t*=2.13, p=0.025
3	TOUCHED (grandmother: girl)	.583	*t*=2.30, p=0.018
4	PULLED (father: child)	.546	*t*=1.17, p=0.13
5	HIT (truck: ball)	.565	*t*=2.43, p=0.014
6	SMACKED (door: branch)	.562	*t*=3.43, p=0.0019

Table S2. Post-hoc analysis of right posterior insula ROIs discovered by the wholebrain searchlight analysis. Classification performance is broken down by mirror-image proposition-pair.

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

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

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

To evaluate the models, we trained the SVR model on N-2 trials, and attempted to predict the frequencies of two held-out observations. The absolute value of the difference between the target frequencies and those predicted by the SVR model was compared for both the correct mapping and the incorrect mapping. If the sum of the two correct mappings had a lower absolute difference than the sum of the two incorrect mappings, the model was determined to have been correct. This procedure was repeated using different frequency metrics, and the results are shown in Supplementary Figure 2. We consistently found no information about any frequency statistics to be available in lmSTC. It is therefore unlikely that our results owe to systematic frequency differences between pairs: rather they appear to reflect the structured, semantic content of sentences.

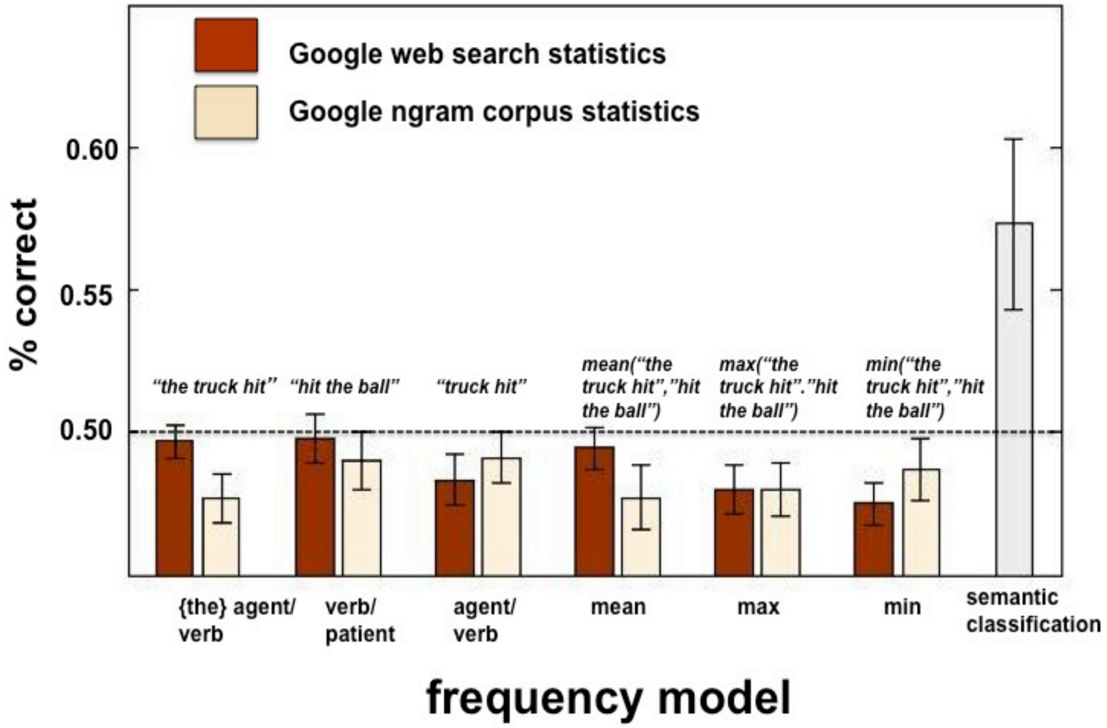

Figure S2 . As a control, we attempted to use lmSTC patterns of activity to predict frequency statistics for NOUN/VERB and VERB/NOUN combinations (e.g., "the truck hit", or "hit the ball"). The patterns of activity did not encode information about any of the frequency statistics tested.

Pattern-Based Effective Connectivity (PBEC) analyses

Background and Motivation. The identification of representational content in the human brain has benefited in recent years from the development of Multi-Voxel Pattern Analysis (MVPA) (25, 38-40). Rather than asking if the magnitude of the BOLD response in a single-voxel or brain region is predicted by the presence or absence of a psychological operation, researchers now routinely pool information across sets of voxels to ask whether and where distributed patterns of activity track variation in psychological content.

MVPA has been productively applied to domains lacking the differential engagement of psychological processes expected to generate uniformly greater neural activity over a brain region (when sampled at the spatial resolution available to contemporary neuroimaging), but in which some representational content nevertheless varies over time (40-42). While this ability makes MVPA particularly well-suited to our current aims, its increased power is not completely without cost (43), as its heightened detection sensitivity makes it more susceptible to subtle confounds. It is therefore particularly important to establish that information detected by the pattern classifier actually reflects the psychological processes or representations of interest.

To address this potential concern, we employed the following reasoning: if the patterns identified reflect neural representations of psychological content, then one would expect the pattern instantiated on a given trial to modulate downstream responses that depend on that content. Given that this logic is broadly consistent with the logic underlying traditional effective connectivity analyses (44, 45), we call the present analysis a *pattern-based effective connectivity* (PBEC) analysis. As with conventional effective connectivity analyses, the aim is to establish the functional *influence* of one neuronal population on another.

Several groups have recently integrated MVPA and functional connectivity analyses, devising ways to determine whether various neural structures share similar representational profiles (46), and whether patterns in one region correlate with univariate responses (47) and patterns (48) elsewhere in the brain. The present analysis extends this integration of MVPA and connectivity analyses to model cases in which the patterns of activity in one region are thought to *drive* the functional state of another using mediation analyses (27). Such tests are widely used in the social sciences (27) and have been previously applied to fMRI data (49, 50). Tests for mediation assess whether the effect of a predictor variable on an outcome variable is either partly or wholly carried by an intervening, or mediating, variable.

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

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

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

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

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

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

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

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

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

where $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are the vectors of class-specific mean voxel intensities across lmSTC. $\boldsymbol{m_1}$ is the affectively salient ("the grandfather kicked the baby") mean vector and $\boldsymbol{m_2}$ the affectively neutral ("the baby kicked the grandfather") mean vector. Σ^{-1} was a shrinkage estimate of the population covariance matrix shared between all stimulus classes. This shrinkage estimate has been shown to be a better estimator of the population statistic than the sample covariance in cases where the ratio of observations to predictors is unfavorable (35, 51, 52), and to perform well with classification of fmri data (33).

The constant term was determined as

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

The weight vector *w* determines the direction through the feature space, while the constant term determines the location of the hyperplane relative to the origin (*56*). Here, the $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ terms for each subject were the means of the respective class along the projection for the remaining 15 subjects. We found these group-level mean estimates to yield more reliable predictions in the connectivity analysis than using an individual subject's data (t(24)=3.93, *p*=0.001 vs. t(24)=1.28, *p*=0.22). These results survive correction for multiple comparisons for these two ways of obtaining w_0.

Finally, the signed distance of an observation from the hyperplane was computed as in (32):

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

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

This distance provides a trial-by-trial measure of the representational content of lmSTC, which can be used as a predictor variable in subsequent connectivity analyses. In the present case, given that the affectively significant propositions occupied the positive region of the space, and the affectively neutral regions occupied the negative region of the space, we predict a positive statistical relationship between trial-by-trial signed distance in lmSTC and the magnitude of the mean signal level across the left amygdala ROI (See Figure 1 of the main text).

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

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

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

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

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

The heat map in Supplementary Figure 3 visualizes the average regression performance for various sigmoids by averaging the search results across 16 cross-validation iterations (one holding each subject out). These search results provide information about the relationship between the pattern in lmSTC and the amygdala's response. For simplicity, if we conceptualize the amygdala response as a binary variable, the center parameter of the sigmoid defines the point at which the probability of a response is equally likely to the probability of a nonresponse. We see from the heat map in Supplementary Figure 3 that the optimal center of the sigmoid is shifted to the right of the classification boundary, in the positive region of the space. The observed positive shift of the center parameter relative to the hyperplane is explicable under the assumption that the probability of the amygdala's *not responding* given that the stimulus is "the grandfather kicked the baby," is greater than the probability of the amygdala's responding given that the stimulus was "the baby kicked the grandfather." This assumption is reasonable given that each proposition is encountered repeatedly over the course of the experiment, potentially attenuating the subject's affective responses over repeated presentations. This would make failures to respond to the affectively salient proposition more likely than "false positives" of the amygdala to the affectively neutral proposition, leading to a positive shift of the optimal center (as the point of equi-probability) for the sigmoid relative to the classification hyperplane. This method may hold more general promise for testing different quantitative models of the functional dependence between brain regions.

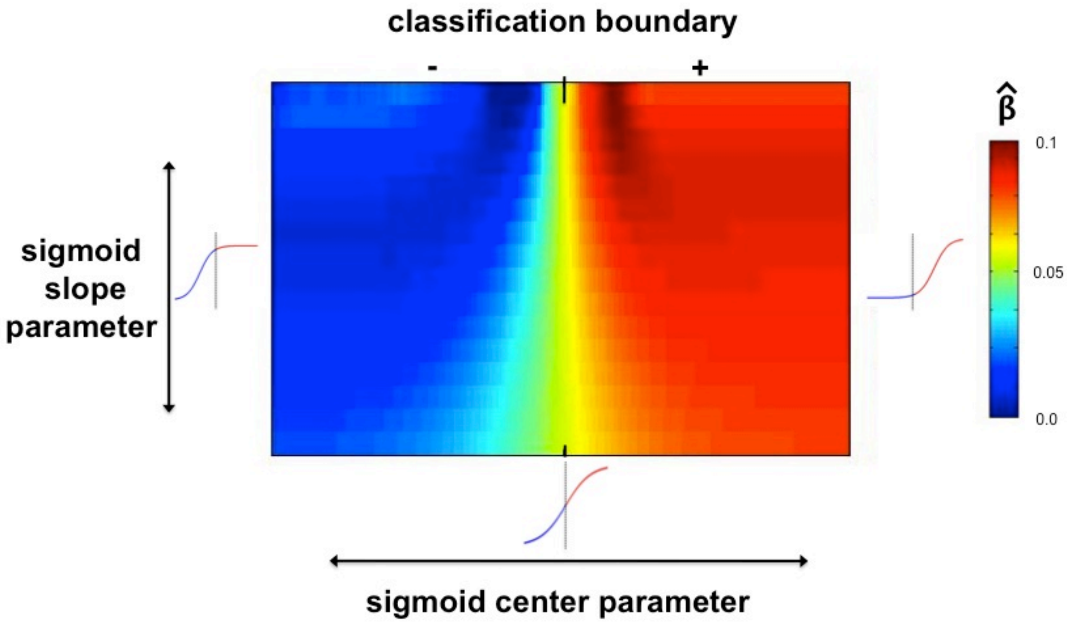

Figure S3. Heat map visualizing variation in the prediction of amygdala activity as a function of different sigmoidal transformations of the signed distance of an observation from the lmSTC classification boundary. Colors correspond to different beta values for this regression averaged across 16 cross-validation iterations. The X-axis represents different values of the sigmoid's "center." The Y-axis represents different slopes, with steeper slopes located at the top of the graph. Together, these two parameters yield different shapes, three of which are shown at the appropriate locations in the space for reference. Vertical bars through the reference sigmoids correspond to the location of the classification boundary, as does the bar at the top along the X-axis. The '+' signifies the positive side of the classification hyperplane (the 'grandfather kicked the baby' side), while the '-' signifies that negative side of the hyperplane. We see better prediction when the sigmoid is centered at or to the right of the hyperplane. The best prediction is obtained by a function with a steep slope, centered slightly to the right of the hyperplane. (See Figure 1B of the main text.) This demonstrates that patterns corresponding to "the grandfather kicked the baby" are

more likely to elicit an amygdala response than their mirror-image, as explained in the main text. However, the rightward-shift of the best sigmoid, past the classification boundary, also suggests that the likelihood of *no* amygdala response, given the “grandfather kicked the baby” pattern in lmSTC was greater than the likelihood of an amygdala response, given the “baby kicked the grandfather” pattern. More generally, this demonstrates that the point of indifference for classification need not coincide with the point of indifference for predicting a response elsewhere in the brain. This approach may be useful for empirically testing subtly different quantitative functions relating information in one brain area to information or activation in other.

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

The statistic is computed as

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

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

As reported in the main text, we found this indirect effect to be statistically significant (z=2.47, *p*=0.013). Alternative nonparametric techniques for estimating confidence intervals for the indirect effect, including bootstrapping (53) and Monte Carlo simulation (54) yielded results comparable to the classic, parametric Sobel test (all *p*<0.05 for the indirect effect). The above analyses were repeated with the right posterior insula/extreme capsule ROI as the mediator. In contrast to lmSTC, all assessments of the pattern of activity in the right insula ROI as the mediating variable were non-significant (*p*>0.15).

Experiment 2: Supporting Information

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

Data Acquisition and Preprocessing. Experiment 2 was conducted using the same facilities and equipment as Experiment 1. However, for Experiment 2, we acquired partial-coverage functional data at higher spatial resolution, using 1.5 mm^3 voxels (FOV = 192mm,TR = 2500 ms, TE=32 ms, 6/8 partial Fourier encoding, Flip Angle = 90°). We acquired 26 slices parallel to the anterior commissure, centered on the superior temporal lobes. While the exact area varied from subject to subject, this ensured that the superior temporal gyri, superior temporal sulci, and mid-portion along the anterior-posterior axis of the middle temporal gyri were covered, bilaterally. A mask generated by the searchlight results of Experiment 1 was then applied to lmSTC prior to analyses, as described below. Experiment 2 employed the same preprocessing operations as Experiment 1, with the additional application of a 1.5mm FHWM smoothing kernel prior to classification analyses.

Experimental Procedure. The sentences for Experiment 2 were generated using four nouns ("man", "girl", "dog", "cat") and five transitive verbs ("chased", "scratched", "blocked", "approached", "bumped") to create every possible agent-verb-patient combination, with the exception of combinations using the same noun twice (e.g. "the dog chased the dog") yielding 60 (4x5x3) unique propositions. These particular verbs were chosen because they permit plausible agent-verb-patient combinations using the above nouns, and are comparable in their frequency of occurrence.

Experiment 2 consisted of six scan runs. Each proposition was presented once per run, and six times in total. Whether a proposition was presented in the active or passive voice on a given trial was randomly determined. As in Experiment 1, sentences were visually presented for 2.5 seconds followed by 7.5 seconds of fixation. A comprehension question was presented following the fixation period on 1/3 of the trials. These questions were of the form "did the dog chase something?" or "was the dog chased by something?" and 50% had affirmative answers.

General Searchlight Procedure. All searchlight analyses for Experiment 2 were confined to the lmSTC. The searched area was formed by dilating the group-level lmSTC ROI discovered in Experiment 1 6mm so as to encompass all of the mid and posterior regions of the left superior temporal gyrus, superior temporal sulcus, and middle temporal gyrus. The resulting ROI contained 6882 1.5 mm^3 voxels, (center, -54, 23, 3). Figure 2b of the main text shows the extent of the searched region for Experiment 2. This mask was warped from Talairach space to each subject's native space, and all classification analyses were conducted in the subject's native space. As in Experiment 1, all searchlight analyses were implemented in the Searchmight

Toolbox (33) and used a linear classifier with a shrinkage estimate of the covariance matrix. Local voxel-neighborhoods were defined using a 3mm (two voxel) radius within the lmSTC mask, entailing that non-edge neighborhoods again contained 124 voxels.

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

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

Verb Decoding Procedure*.* We performed an additional searchlight analysis probing lmSTC for sub-regions that contained information about the trial-by-trial identity of the sentence's verb. This task required a five-way decision regarding the identity of the verb ("chased", "scratched", "blocked", "approached", "bumped"). Here the cross-validation folds were defined over the 6 scanning runs. The classifiers were trained using data from 5/6 runs, and were then asked to identify the verb present for trials from the remaining run. The verb classifiers were thus only required to generalize to new tokens of previously encountered combinations, rather than wholly new combinations. Individual accuracy maps were averaged as in prior analyses, smoothed with a 3mm FWHM kernel, warped to Talairach space, and submitted to a one-tailed *t*-test against 0.2, as chance performance for the five-way verb classification was 20%. Corrected *p* values were obtained in the same manner as in the above searchlight analyses. Results are visualized in Figure S4., and more information is provided about the analysis and results in Supplementary Table 3.

Experiment 2 searchlight results.

identification problem	center coordinates (Talairach)	# voxels	Volume (mm^3)	smoothness of noise map (XYZ, mm)	corrected P(Data|Noise)
Agent	-46, -18, 1	67	100.5	4.27, 5.72, 4.93	p<0.01
	-57, -37, 7	61	91.5		p<0.02
Patient	-57, -10, 2	181	271.5	5.13, 5.6, 5.47	p<0.0001
Verb	-61, -15, 2	60	90	4.59, 6.17, 5.08	p<0.025
	-55, -49, 5	433	649.5		p<0.0001

Table S3. Searchlight results for Experiment 2 listing significant clusters within lmSTC. All searchlight analyses used a voxelwise threshold of p<0.005, and a corrected threshold of p<0.05 to identify significant clusters.

Post-Hoc ROI Analyses

Representational Specificity of Sub-Regions. For all post-hoc analyses, we used a leave-one-subject-out cross-validation procedure to localize regions of interest without biasing the analyses. Specifically, we iteratively conducted group-level *t*-tests on the search maps for 24/25 subjects to identify statistically significant clusters of informative voxels with each subject's data removed. All such regions were localized using a voxelwise threshold of p<0.005 and a minimum cluster size of 50 voxels, unless otherwise specified. These significant voxel-clusters defined the exact ROI for the held-out subject for the post-hoc test of interest. For any post-hoc analysis, the exact ROI queried could thus vary slightly from subject to subject, though the number and general location of ROIs was stable across all cross-validation iterations of all analyses.

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

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

To directly compare such "role-neutral" detection functions to the "role-based" identification functions described throughout, we first converted the accuracies obtained by the searchlight procedure to probabilities using a binomial distribution. Unlike previous comparisons, this conversion was necessary given that "chance" levels differed between the two classification functions: chance performance for "role-neutral" classifiers was 50%, whereas chance performance for "role-based" classifiers was 25%. We localized the agent and patient regions using the leave-one-subject-out method described above, and averaged the "role-neutral" and "role-based" *p* values across the regions. We then compared these two sets of average *p* values by region using a *t*-test.

Averaging across the two agent regions, we found them to be significantly better at identifying the agent than determining whether a particular noun was present (t(2.57), *p*=.017). When these two agent regions were treated separately, accuracy in agent identification was significantly better than noun detection in the posterior agent region (t(24) = 2.20, *p*=0.037), and was marginally better in the anterior agent region t(24)=1.92, *p*=0.066), As expected, the patient region was significantly better at identifying the patient than at role-neutral noun detection (t(24) =2.65, *p*=0.014).

These findings indicate that our principal verb-general searchlight results are not simply detecting the semantic content of the nouns without respect to role. Instead, they encode information about the identity of the agent and patient across verbs.

Medial-Lateral Agent/Patient Topography of anterior lmSTC*.* Figure 3c in the main text visualizes the representational content of anterior lmSTC moving along the medial-lateral axis. Other researchers have used similar analyses to assess the representational topography of objects in the ventral temporal cortices (Mahon et al., 2009; Connolly et al. 2012) . In performing these analyses, it was important to first localize the anterior lmSTC in a way that is unbiased with respect to its role preferences. To do this, we asked where in lmSTC a classifier could reliably tell, for a given noun (e.g. MAN), whether that concept was the agent or patient across trials. We conducted four such analyses, one for each noun, each time focusing only on trials in which the target noun was either the agent or patient. This analysis *jointly* localizes the set of sub-regions that encode the identity of the agent along with those that encode the identity of the patient. Because, for a given noun, being the agent is perfectly correlated with *not being* the patient and vice versa (in this analysis, though not in the experiment more broadly), both agent and patient regions should be identified by this analysis. Critically, this analysis reveals nothing about which type of information (agent/patient) is encoded in a particular sub-region, making it a suitably unbiased localizer for present purposes.

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

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

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

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

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

We found that the neural representations in both the patient ROI and the posterior agent ROI automatically generalize across active and passive sentence forms ($t(24) = 2.10$, p=0.023, one-tailed: $t(24)$=1.83, p=0.039), but that the anterior agent ROI patterns did not ($t(24)$=1.10, p=0.14). While this difference may signal a functional difference between the two agent ROIs, it is important to note that this training procedure uses 50% of the data used by the searchlight analyses (288 trials vs. 144 trials) making it difficult to interpret this null result conclusively.

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

To determine whether the neural representations generalize to all verbs used, we performed a searchlight analysis in which local classifiers were (1) trained to make the same four-way agent/patient identification decisions described above using data generated by the target verb (20% of the data) and asked to generalize to the remaining four verbs (80 % of the data), and (2) trained using the remaining four verbs (80% of the data) and asked to generalize to the target verb (20% of the data). The results of (1) and (2) were then averaged to provide a measure of how well the patterns of activity corresponding to noun/role combinations in that verb-context generalize to the other four verb-contexts. This analysis produced 5 separate search maps of lmSTC for agent/patient identification in generalizing to each of the five verbs. We iteratively localized clusters containing information for 4/5 verbs using these search maps and a liberal threshold (p<0.05 voxelwise, k=50), and asked whether the average classification accuracy in generalizing to the fifth verb in the identified region was significantly greater than chance. Supplementary

Table 4 shows the results of these analyses for the three ROIs, by generalization to each verb.

We then performed a similar analysis examining classification performance, broken down by the nouns occupying these roles. Here, we performed separate searchlight analyses for each of the 6 possible pair-wise noun discriminations (e.g., discriminating "man as agent" vs. "girl as agent"), for both the agent and patient roles. The ROIs were again iteratively localized using the results of 5/6 pair-wise classifications and a liberal threshold (*p*<0.05 voxelwise, k=50), and the searchlight results of 6th pair were averaged across the resulting ROI. The results are presented for each pair-wise comparison in Supplementary Table 4.

(B). Generalization by noun pair

Noun Pair Discrimination	patient	Ant-agent	Post- agent
Man/Girl	1.85, p=0.039	2.68, p=0.0065	3.21, p=0.0019
Man/Dog	1.23, p=0.115	1.47, p=0.077	2.10, p=0.023
Man/Cat	2.90, p=0.004	2.22, p=0.018	2.02, p=0.027
Girl/Dog	1.20,p=0.121	1.08, p=0.145	2.62, p=0.0075
Girl/Cat	3.53, p=0.0009	3.2, p=0.0019	1.56, p=0.066
Dog/Cat	1.71, p=0.05	1.20, p=0.121	1.06, p=0.15

Supplementary Table 4. Post-hoc analyses of agent and patient identification performance within their corresponding ROIs when generalizing to (a) each verb context and (b) new pair-wise noun discriminations. Values are *t*-statistics, with one-tailed *p* values against chance.

Experiment 3 Supporting Information

Data Acquisition and Preprocessing. Experiment 3 was conducted using the same facilities and equipment as Experiments 1 and 2. A high-resolution structural scan (1mm^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 = 192mm,TR = 3500 ms, TE=28 ms, Flip Angle = 90°). We used parallel imaging (iPAT 2) to obtain whole-brain coverage with 2x2x2 mm voxels. We obtained whole-brain data in order to assess lmSTC's functional connectivity to various cortical regions. Those analyses are not presented in this thesis. For the analysis presented here, we applied a mask generated by the dilating the anterior agent and patient searchlight regions from Experiment 2 by 8 mm prior to analysis.

Experiment 2 employed the same preprocessing operations as Experiments 1 and 2. The data were not smoothed prior to classification analyses.

Experimental Procedure. Sentences were constructed from a menu of 6 nouns and 8 transitive verbs, to create every possible aRb proposition, excluding propositions in which the same noun occupied both roles (e.g., "the goose approached the goose"). Participants thus read 6x8x5=240 unique propositions, each presented once. This contrasts with Experiments 1 and 2, in which each proposition was presented multiple times. Whether a proposition was presented in the active or passive voice was randomly determined, and 50% of the propositions were presented in each voice.

We used 6 monosyllabic English nouns that refer to animals ('moose', 'cow', 'hog', 'crow', 'goose', 'hawk). These nouns were chosen because their phonological similarity relationships are distinct from their semantic similarity relationships. Semantically, these nouns most 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 ('hawk' in the bird class). We were guided by Mueller et al.'s (2003) PSIMETRICA model of phonological similarity in choosing nouns with systematic phonological similarity relationships. On this characterization, 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", which determines our intuitive sense for whether two syllables rhyme. Our stimuli consist of three pairs of phonologically similar nouns. Each phonologically similar pair is similar along 2/3 of Mueller et al.'s, dimensions: similar onset and nucleus {'hog','hawk'}, similar onset and coda {'crow','cow'}, or similar nucleus and coda {'goose','moose'}.

To verify that participants indeed treated the expected stimuli as simiarl, and independent group of participants (n=50) on Amazon Mechanical Turk performed pairwise similarity ratings on these items, rating both how similar the words sound and how similar the animals are. They rated each pair using a sliding scale that ranged from 0-100. Qualitatively, zero was labeled to indicate "not similar at all", 50 to indicate "somewhat similar", and 100 to indicate "very similar". All participants performed both the semantic and phonological ratings; the order of which was randomized. These ratings confirmed that the stimuli had the similarity structure we expected (See Figure 5). We used these participants' ratings to define the similarity spaces with which we would model the similarity structure of other participants' brain responses.

We used 4 verbs that conveyed some aspect of a mental state, which we will, following others, refer to as "psych verbs", and 4 verbs that did not explicitly convey mental state information, but were characterized primarily by different manners of motion of the agent with respect to the patient/theme. These were 'chased', 'approached', 'passed', and 'attacked'. The four psych verbs broke down into two pairs, the members of which were similar both semantically and syntactically: ('surprised', 'frightened'), and ('noticed', 'detected').

For the first 4 scan runs, participants read these nouns sequentially, each presented by itself independent of any sentence context. On a given trial, one noun would be presented visually to the subject for 3.5 seconds, followed by 7 seconds of fixation. After 7 seconds, participants could be asked a simple general knowledge question about the animal (e.g., ("does it lay eggs?"), or a simple question about the word's phonology ("does it rhyme with 'loose'?"). These questions were asked on % of the trials. 50% of the questions were general knowledge questions, and 50% were questions about phonology. On trials in which there was no question, the stimulus presentation simply advanced to the next noun following the 7-second fixation period. Each noun was presented 6 times per run, and 24 times total over the course of the four runs.

For the last 6 scan runs, participants read the seconds generated by combining the nouns and verbs in the manner described above. On a given trial, one sentence would be presented visually to the subject for 3.5 seconds, followed by 7 seconds of fixation. After 7 seconds, participants could be asked a question about who did what to whom in the sentence they had just read. These questions were asked on % of the trials. 50% of the questions targeted the underlying subject (e.g., "did the moose approach something?"), and 50% targeted the underlying object ("was the moose approached by something?"). On trials in which no question was asked, stimulus presentation simply advanced to the next noun following the 7-second fixation period.

Searchlight analyses.

All searchlight analyses for Experiment 3 were confined to the lmSTC. The searched area was formed by dilating the group-level agent and patient ROIs discovered in Experiment 2 8mm so as to encompass all of the mid and posterior regions of the left superior temporal gyrus, superior temporal sulcus, and middle temporal gyrus. The resulting ROI contained 599 2mm^3 voxels. This mask was warped from Talairach space to each subject's native space, and all classification analyses were conducted in the subject's native space. As in Experiment 1, all searchlight analyses were implemented in the Searchmight Toolbox (Pereira & Botvinick, 2011) and used a linear classifier with a shrinkage estimate of the covariance matrix. Local voxel-neighborhoods were defined using a 2mm (one voxel) radius within the lmSTC mask, entailing that non-edge neighborhoods again contained 27 voxels.

For all classification analyses, searchlights iteratively selected data from local voxel-neighborhoods to make six-way decisions regarding the noun occupying the agent or patient role on a given trial (moose? goose? cow? crow? hog? hawk?). As in Experiments 1 and 2, active and passive versions of the same proposition were considered identical for the purposes of these analyses. We performed two types of classification analyses. The first mirrored the principal analyses of Experiment 2. Here, we established a target function (e.g., the agent across motion verbs), trained on N-1 verbs using that function, and tested the classifier's ability to identify the value of that variable for the held-out verb. We performed this analysis for the following functions: agent, the patient, the stimulus, the experiencer, [agent OR stimulus], [patient OR stimulus], [agent OR experiencer], [patient OR experiencer]. The second type of classification required the classifier to generalize automatically to a new variable. For example, we trained the classifier to identify the value of the experiencer variable using only experiencer-subject verbs, and tested its ability to identify experiencer-object verbs, vice versa, and then averaged the two results. We performed this cross-training/testing method for the experiencer across syntactic constructions, the stimulus across syntactic constructions, agent TO stimulus, agent TO experiencer, patient TO stimulus, patient TO experiencer (and vice versa for these cross-role classifications). For these cross-training analyses, we trained on one category and tested on the second, then trained on the second and tested on the first. We then averaged these two iterations to obtain the final classification accuracy.

For all analyses, classification accuracies for each subject were averaged across cross-validation folds, and the mean accuracy was assigned to the center voxel of the search volume. These individual-level accuracy maps were then smoothed with a 2mm FWHM kernel, warped to Talairach space, and the group of subjects' maps was submitted to a directional one-sample *t*-test against 0.16667 to determine whether any regions reliably encoded information about the identity of the agent or patient across subjects. We used Monte Carlo simulation to determine the probability of obtaining significant clusters given that the data contained only noise. We used a voxelwise threshold of $p<0.05$, k=30, and a corrected threshold of $p<0.05$ to identify regions exhibiting statistically significant effects. Statistical correction was performed within searchlight analyses, correcting for multiple search neighborhoods, but not across multiple searchlights. As in Experiment 1, we estimated the smoothness of the data by conducting the same classification and aggregation procedures with randomly permuted labels. Post-hoc analysis were done as described within group-level ROIs.

Post-hoc Analyses

For all post-hoc analyses, we used a leave-one-subject-out cross-validation procedure to localize regions of interest without biasing the analyses, as in Experiment 2. Specifically, we iteratively conducted group-level *t*-tests on the search maps for 39/40 subjects to identify statistically significant clusters of informative voxels with each subject's data removed. All such regions were localized using a voxelwise threshold of $p<0.05$ and a minimum cluster size of 25 voxels. These significant voxel-clusters defined the exact ROI for the held-out subject for the post-hoc test of interest. For any post-hoc analysis, the exact ROI queried could thus vary slightly from subject to subject, though the number and general location of ROIs was stable across all cross-validation iterations of all analyses. We performed repeated-measures ANOVAs to test for a sub-region by role interaction for the anterior agent and patient regions, the stimulus and experiencer sub-regions, and the agent/stimulus and patient/experiencer sub-regions.