EVENT STRUCTURE IN VISION AND LANGUAGE

Alon Hafri

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Co-Supervisor of Dissertation

_____

John C. Trueswell, Ph.D.

Professor of Psychology

Co-Supervisor of Dissertation

_____

Russell A. Epstein, Ph.D.

Professor of Psychology

Graduate Group Chairperson

_____

Sara R. Jaffee, Ph.D.

Professor of Psychology

Dissertation Committee

Nicole Rust, Ph.D.

Associate Professor of Psychology

Konrad Kording, Ph.D.

Penn Integrates Knowledge Professor

ACKNOWLEDGMENT

This work was made possible by a range of financial and institutional support. For supporting the research, I am grateful for funds from the Department of Psychology, from the Penn Center for Magnetic Resonance Imaging & Spectroscopy, and from grants awarded to my advisors from the National Science Foundation and National Institutes of Health. For supporting me, I am grateful for a graduate research fellowship and IGERT training grant from the National Science Foundation, and a Penn Vision Training Grant from the National Institutes of Health.

First, I owe a great deal of gratitude to my dissertation committee, Nicole Rust and Konrad Kording, for their cutting questions and constructive comments. Their presence at my meetings has been an honor and a privilege, and I hope that the quality of the work in this thesis meets and exceeds their expectations.

I am also very lucky to have been surrounded by such kind and brilliant peers (and eccentric, in the best way): in Russell Epstein's lab, Josh Julian, Alexandra Keinath, Steve Marchette, Mick Bonner, Jack Ryan, Zhengang Lu, Michael Peer, Rachel Metzgar, and Nick Dewind; in John Trueswell's lab, Lucia Pozzan, Felix Wang, Alex de Carvalho, Victor Gomes, and Monica Do; and Chris Angeloni, my science and brew partner. It has been quite a ride… the late-night science debates intermixed with games, soup-making, and recorder + ukulele jams ("Night after night, we make the same mistakes!").

Several others have had a great influence on my thinking and approach to science. Brent Strickland, with whom I collaborated on Chapter 2 among other projects, has been a great inspiration to me. His creative and bold approach to science inspires me to approach it in the same way. Lila Gleitman's decades of research on language and thought have had a profound effect on what questions I consider are the deep and interesting ones to ask. I have also enjoyed our many intellectual conversations about how many people it takes to tango (quite a question, I'm sure she would agree).

Over the last six years, I have been astoundingly lucky to work with a true advisory dream-team, John Trueswell and Russell Epstein. They are the primary reason I chose to stay at Penn for graduate school instead of going off to a galaxy far, far away. Russell has been incredibly supportive of my work, even though my research path has moved far from spatial navigation per se. In fact, this intellectual distance has proved most

valuable, as Russell asks really hard questions about the nature of our projects, both theoretical and methodological. I have known John for about ten years now (!), and over this time my appreciation for his approach to science has only grown. His endless confidence in my ability as a scientist has kept me going, even when experiments have sometimes not gone as planned. Together, Russell and John have allowed me to pursue my own ideas while remaining supremely accessible for feedback and advice. They are major sources of inspiration for how to be both a productive scientist and encouraging mentor, and I am no doubt a better scientist because of them.

Finally, my family and friends have been crucial to my sanity over these years, offering endless outlets for performing music, brewing beer, and making many gallons of soup. Gabi the cat has also provided distractions of fuzziness, as necessary. Unsurpassed in all respects of course is Lisa Rothstein, my partner. Lisa has supported me with her weirdness, kindness, funniness… the list goes on. Did I say weirdness? She has made these last six years richer and more joyful in all respects, academic and otherwise. I would not have been able to do this without her.

Published chapters

The findings from Chapter 2 have been published as: Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52. http://doi.org/https://doi.org/10.1016/j.cognition.2018.02.011

The findings from Chapter 3 have been published as: Hafri, A., Trueswell, J. C., & Epstein, R. A. (2017). Neural representations of observed actions generalize across static and dynamic visual input. *The Journal of Neuroscience*, *37*(11), 2496–16. http://doi.org/10.1523/JNEUROSCI.2496-16.2017

Data availability

For Chapter 2: All raw data is publicly available with the Open Science Framework, at the following link: https://osf.io/uewfn.

For Chapter 4: All de-identified fMRI data will be made publicly available on the CRCNS web site upon publication (http://crcns.org). Stimuli and model feature data will be made publicly available at the Open Science Framework (https://osf.io), including

human ratings collected for implementing the Binder model and VerbNet verb annotations made by the first author. Relevant code for extracting VerbNet features and for conducting analyses will be made available on github (https://github.com/ahafri).

ABSTRACT

EVENT STRUCTURE IN VISION AND LANGUAGE

Alon Hafri
John C. Trueswell
Russell A. Epstein

Our visual experience is surprisingly rich: We do not only see low-level properties such as colors or contours; we also see *events*, or *what is happening*. Within linguistics, the examination of how we talk about events suggests that relatively abstract elements exist in the mind which pertain to the relational structure of events, including general thematic roles (e.g., Agent), Causation, Motion, and Transfer. For example, "Alex gave Jesse flowers" and "Jesse gave Alex flowers" both refer to an event of transfer, with the directionality of the transfer having different social consequences. The goal of the present research is to examine the extent to which abstract event information of this sort (event structure) is generated in visual perceptual processing. Do we *perceive* this information, just as we do with more 'traditional' visual properties like color and shape? In the first study (Chapter 2), I used a novel behavioral paradigm to show that event roles – who is acting on whom – are rapidly and automatically extracted from visual scenes, even when participants are engaged in an orthogonal task, such as color or gender identification. In the second study (Chapter 3), I provided functional magnetic resonance (fMRI) evidence for commonality in content between neural representations elicited by static snapshots of actions and by full, dynamic action sequences. These two studies suggest that relatively abstract representations of events are spontaneously extracted from sparse visual information. In the final study (Chapter 4), I return to language, the initial inspiration for my investigations of events in vision. Here I test the hypothesis that the human brain represents verbs in part via their associated event structures. Using a model of verbs based on event-structure semantic features (e.g., Cause, Motion, Transfer), it was possible to successfully predict fMRI responses in language-selective brain regions as people engaged in real-time comprehension of naturalistic speech. Taken together, my research reveals that in both perception and

language, the mind rapidly constructs a representation of the world that includes events with relational structure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Figure 1.1
What's happening? How do we know?

Our visual experience is surprisingly rich: We see not only colors or contours, objects or scenes, but also *what's happening*. And, indeed, it seems we cannot help but do so. For example, Figure 1.1 could in principle be described as "Some people on a field", but instead you would likely remark, "The red player savagely bit the blue player's arm!". Recognizing what is happening "out there" is central to our everyday experience in a physical and social world, yet the nature of this recognition process is unclear.

Decades of empirical work has had success in addressing one aspect of the problem: the visual processes and neural systems involved in recognizing categories of movement patterns, such as walking or running (Giese & Poggio, 2003; Lange & Lappe, 2006), or of hand-object interactions, such as grasping (Fleischer, Caggiano, Thier, & Giese, 2013; Rizzolatti & Sinigaglia, 2010). Yet consider again the soccer star biting his opponent. From this it is clear that we do not merely recognize patterns of motion (in fact, Figure 1.1 is devoid of explicit motion signals). Instead we recognize events, with a rich internal structure: the player in red bit the player in blue.

Event structure specifies the kinds of relationships between entities and the environment, and the roles that each entity plays in the event. In Figure 1.1, the red player is an Agent (the entity who acts) and the blue player is a Patient (the entity who is acted on or undergoes a change of state). Recognizing this structure allows us to make rich inferences about the dispositions of the individuals (Hamlin, Wynn, & Bloom, 2007), or who might deserve blame (De Freitas & Alvarez, 2018) — all of which are fundamental for interacting with individuals and for reasoning about the world. Work in linguistics offers a rich set of theoretical predictions about what the components of event structure might be. Strikingly, however, there has not been systematic investigation of the perceptual processing of event structure in its own right.

What is the nature of event representations in the mind? Do they merely arise in our explicit, effortful judgments about our environment, perhaps based on reasoning about the relative locations and poses of entities? Or might they have a basis in more primitive

and foundational mental processes, such as the rapid, automatic processes of visual perception and attention? This latter possibility is in line with growing evidence that the visual system itself traffics in high-level properties like causality (Rolfs, Dambacher, & Cavanagh, 2013), animacy (van Buren, Uddenberg, & Scholl, 2015), and balance (Firestone & Keil, 2016).

The representation of event structure elicited by visual and linguistic input is the central topic of this thesis. In the following chapters, we employ techniques to predict behavior and brain activity as people engage in tasks involving event perception and language comprehension. Across three studies, we explore the automaticity of event structure elicited by these two modalities. In each study, we test the content and generality of these representations. Taken together, this research program lends support to the hypothesis that the perceptual system traffics in high-level representations of events that have internal structure of a very particular sort. In particular, the elements of this event structure correspond to those made explicit by theories about the relationship between the structure of language and the structure of events.

Below, we briefly review relevant linguistic and developmental literature that indicates the kinds of representations we might expect to be afforded by the visual system. We then discuss approaches taken thus far to investigate the representations of events afforded by high-level vision. Finally, we preview the studies conducted in this thesis.

1.1. Approaches to event structure in the mind: Linguistic and developmental evidence

What is an event category? As with the correspondence between nouns and object categories, the verbs and verb phrases used in a language can give us a preliminary idea of the kinds of event categories that humans conceptualize. For example, we conceive of *biting* and *hugging, kissing* and *kicking*. Delving deeper, a moment's thought reveals that each category involves a particular kind of spatiotemporal occurrence: for biting, a person or animal's teeth must enclose something. Simply reaching one's mouth toward something does not suffice. This demonstrates the intuition that a single verb can refer to a combination of several components; but what are these components?

Before contemporary cognitive science traditions, philosophers recognized the

centrality of certain conceptual components of events to our understanding of the world (e.g., causality; Hume, 1739). More recently, the efforts of cognitive psychologists to account for patterns of linguistic structure within and across languages has become a fruitful source for theories about what components of events the mind represents. For a simple example, consider the following sentences, where the noun phrases before and after the verb are underlined:

1) The red player ran.
2) The red player bit his opponent.
3) The referee gave the red player a penalty.

Contrast the previous examples with the following, which are odd (??) or ungrammatical (*):

4) ?? The red player ran his opponent a penalty.
5) ?? The red player bit.
6) * The referee gave the red player.

Notice that the number of noun phrases is dictated by the nature of the events referred to: running requires one entity, biting two, and giving three. It turns out that these patterns are quite consistent, holding across verbs that refer to events that differ drastically in content (Fisher, Gleitman, & Gleitman, 1991; Jackendoff, 1990; Talmy, 2000). Take the below for example, where 7-9 correspond to 1-3, respectively:

7) The TIE fighter flew.
8) Obi-Wan swung his sword.
9) Luke told Leia the news.

More sophisticated analyses of the patterns of linguistic structure within and across languages have uncovered elements of event structure that are quite general in nature. These elements include, among others, Causation, Motion, and Change of State (detailed further in Chapter 4), as well as sets of event roles (detailed further in Chapter 2). Event

roles (also called thematic roles) describe the specific relationship between entities in an event. For example, in *The girl pushes the boy*, *girl* is the Agent (or one who acts), and *boy* is the Patient (or one who is acted upon or changes state). Role information is often (but not always) indicated by the relative positioning of noun phrases in an utterance (or in some languages, by distinct case markings):

10) Luke killed Darth Vader.
11) Darth Vader killed Luke.

There is a wealth of literature suggesting that these event structure elements are not purely linguistic in nature, but rather are fundamental properties of the mind. Both causation and event roles arrive early in development and organize infants' inferences of the physical and social world (Baillargeon et al., 2012). Pre-linguistic infants as young as 6 months are sensitive to the spatiotemporal properties of causation in simple Michottean launching events (Leslie & Keeble, 1987). Furthermore, infants categorize the entities in such events as distinct roles (Causer and Causee). Compelling evidence that roles are a fundamental conceptual component in the mind comes from deaf children's homesign (Feldman, Goldin-Meadow, & Gleitman, 1978; Goldin-Meadow & Feldman, 1977). These children have no exposure to a natural human language, oral or signed, yet in the signs that they spontaneously produce, the positions of different roles (such as Agent and Patient) within their sequences are consistent within and across individuals. This suggests that language does not externally impose structure on the world that humans experience, but rather that the mind is predisposed to categorize entities in different events into common relational categories (e.g., there is a commonality among the properties of the Agent across kicking and pushing events). There is additional evidence that the perceptual system itself is engaged in extracting certain event components such as causation: retinotopic adaptation to simple causal collisions has been observed in adults (Rolfs et al., 2013). Together, these infant and adult studies suggest that that humans parse the visual world into events, starting from early in development.

However, apart from studies on perceptual causality, the extent and role of the visual system in extracting event information is not known. Recognizing specific instances of

event components – including causation, event roles, and changes of state, among others – must be mediated through perceptual input. Below we briefly review literature on what is known about how perceptual and neural systems extract information relevant for identifying events.

1.2. Approaches to events in vision: Motion patterns and object interactions

There are two main lines of work in vision that are relevant for our interest in event recognition.[1] The first is work on perception of patterns of body movements and the stages of visual processing leading to their extraction. This work generally recognizes distinct contributions of body form and motion to the recognition process, which is well captured by the two-stream form/motion model of Giese & Poggio (2003). The form pathway, whose neural locii are regions in lateral and ventral occipitotemporal cortex, contains shape representations of the human body (Downing, Jiang, Shuman, & Kanwisher, 2001; Taylor, Wiggett, & Downing, 2007) that are integrated over short timespans of about 200 ms ("snapshots" of an action; (Singer & Sheinberg, 2010; Vangeneugden et al., 2011)). For motion extraction, it is established that the necessary computations are performed in area hMT+ (the human middle temporal complex; Rust, Mante, Simoncelli, & Movshon, 2006; Salzman, Britten, & Newsome, 1990; Simoncelli, Heegert, & Heeger, 1998). Finally, evidence from both non-human primate electrophysiology and human fMRI work suggests that form and motion information converge in the posterior superior temporal sulcus (pSTS; Grossman & Blake, 2002; Oram & Perrett, 1994). Such motion pattern representations start out as view-dependent, with greater viewpoint-tolerance achieved at later stages (Jellema & Perrett, 2006; Oram & Perrett, 1994). This latter point is important, because to represent a particular motion pattern (e.g., running), the neural population must encode this information similarly across viewpoints. Although such work has been influential in our understanding of the complex visual computations involved in high-level motion

[1] In much of the visual perception literature, recognition of events has been called, variously, action observation and biological motion recognition. Here I use the term *event* as it reflects the diverse set of spatial, temporal, and state changes characterized in language. But of course, *actions* can be considered a subset of events.

processing, it has thus far not engaged with the larger question of how we recognize categories of relationships between entities.

The second relevant line of work gets closer to this question, by focusing on recognition of interacting objects as distinct perceptual units, and the neural systems that support such representations. Green and Hummel (2006) provided behavioral evidence that identification of individual objects from briefly presented scenes is facilitated when those objects are in interactive spatial orientations (e.g., a tea kettle facing a cup), suggesting that there is an interactive effect for processing objects in canonical interactive orientations. Human fMRI and neuropsychological evidence suggests that human lateral occipital complex (LOC) supports representations of interacting objects as distinct from their components (Kim & Biederman, 2011; Kim, Biederman, & Juan, 2011; Roberts & Humphreys, 2010). Perceptually grouping familiar objects for recognition has important implications for how vision encodes familiar scenes. However, it is not clear whether such representations are "event-like". Events are composed of sets of individuals interacting in specific ways, but importantly, we can recognize them "from scratch": we know *biting* when we see it, even if we observe an unfamiliar individual biting an unfortunate object (or person!) in an unfamiliar setting.

1.3. The current approach: Event structure as the target of perceptual processes (Chapters 2 and 3)

Although the aforementioned work suggests that the visual system processes rich information about human movements and object interactions, and does so in a surprisingly general manner, neither of these approaches gets at the richness of event structure. In our initial investigations of event structure in the visual domain (Chapters 2 and 3), we focus our efforts on event roles (e.g., Agent and Patient, Chapter 2) and event categories (e.g., kicking, Chapter 3). Finding evidence for such representations elicited spontaneously by visual input is proof-of-concept that such structure lies within the domain of perception.

First, we devote our efforts to finding evidence that the visual system itself traffics in the structure of events. Here we take inspiration from decades of work in scene and object perception showing rapid and bottom-up activation of object and scene categories from input lasting less than the span of a fixation (Biederman, Blickle, Teitelbaum, &

Klatsky, 1988; Biederman, Mezzanotte, & Rabinowitz, 1982; Oliva & Torralba, 2001; Potter, 1976; Thorpe, Fize, & Marlot, 1996). In previous work (Hafri, Papafragou, & Trueswell, 2013), we used such paradigms to investigate the recognition of event information under brief display. However, all tasks in previous studies on events, including our own, cannot answer questions about the rapidity, automaticity, or generality of representations extracted, as participants gave their responses several seconds after the image was masked.

In Chapter 2, we overcome these issues. We developed a novel scene priming paradigm to ask whether the human visual system rapidly and spontaneously encodes who acted on whom, or the event roles (e.g., *boy hitting girl* is different from *girl hitting boy)*. Participants observed a continuous sequence of two-person scenes and simply had to identify the male or female (or red or blue-shirted person in another version) in each image. Critically, although role was never explicitly mentioned and was irrelevant for the task, we observed a response switching cost: participants responded more slowly when the target's role switched from trial to trial (e.g., the male went from being the Patient to the Agent, or vice-versa). The experiments in this chapter demonstrate that extraction of event structure from visual scenes is rapid and spontaneous. They further demonstrate the generality of event role representations extracted: the effect was observable across many event types (e.g., Agent of kicking and Agent of pushing). This is predicted by a theory of event structure whose components generalize across a wide range of events.

In Chapter 3, we investigate the neural systems that are associated with action recognition, testing a prediction about the kind of visual input necessary to elicit action representations in such regions. If the target of the recognition process is the relationship between entities in a scene rather than the movement patterns associated with an action, we should observe commonality in content between neural representations elicited by full, dynamic action sequences and by static snapshots of actions (of the kind in Figure 1.1). We test this possibility in Chapter 3. Human participants were scanned with fMRI while viewing categories of interactions (e.g., *pulling*) depicted in two visual formats: (1) controlled videos of two interacting actors; and (2) visually varied photographs selected from the internet involving different actors, objects, and settings. Action category was decodable across visual formats in brain areas previously observed to respond to observed actions, including bilateral inferior parietal,

bilateral occipitotemporal, and left premotor cortex. These results suggest that surprisingly abstract representations of actions are elicited from sparse visual information.

1.4. Modeling event structure in the brain (Chapter 4)

In Chapter 4, we circle back to language: The goal of this chapter is to provide evidence that the semantic system in the brain is sensitive to components of event structure during real-time comprehension of language. Although there is a wealth of research in linguistics and development suggesting the existence of such representations (see section 1.1. above), there is surprisingly little neuroscientific evidence for such distinctions. The work that does exist is limited to studies presenting isolate words to individuals (Kemmerer & Gonzalez-Castillo, 2010). The single-word method is one way to isolate the elements of interest (i.e. the semantic components of verbs), which is especially useful since the elements co-vary in verbs (e.g., *kicking* involves both Cause and Motion). However, such an approach may not reflect the richness of semantic experience when hearing language "in the wild", i.e. from naturalistic input. To overcome this limitation, we chose to use a voxel-wise encoding model approach. Encoding models were first implemented in fMRI to study low- and high-level vision (Kay, Naselaris, Prenger, & Gallant, 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Nishimoto et al., 2011) and were then extended to test models of semantics in natural language (Huth, Heer, Griffiths, Theunissen, & Gallant, 2016). Using this approach allows us to achieve a quantitative description of how simultaneously active event features (e.g., Cause and Motion) are encoded in the brain.

Here we use fMRI to test the hypothesis that the human brain represents verbs in natural language in part via the event structures to which they refer. We scanned participants with fMRI as they listened to audiobook excerpts. Using the encoding model approach, we were able to successfully predict fMRI responses to verbs in language-selective regions using a model based on event structure features (e.g., Cause, Motion, State). In additional comparisons with other linguistic models, we confirmed one prediction of lexical semantic theory: that there is a high correspondence between the linguistic structures that a verb takes and its semantic interpretation. These results suggest that properties of semantic structure (e.g. Cause) are encoded spontaneously by

people as they comprehend naturalistic speech. More generally, this modeling approach provides the technical foundation for future tests of hypotheses about the physiological basis of event representation from other modalities of input, including vision.

1.5. Summary of approach

Taken together, this thesis establishes correspondences between the targets of event recognition and the perceptual and neural systems that contribute to the recognition process. We provide evidence for the hypothesis that that the visual system spontaneously extracts abstract representations of events that includes their internal structure. Similar structure is also elicited by linguistic input and is predicted by the linguistic structural patterns therein.

II. ENCODING OF EVENT ROLES FROM VISUAL SCENES IS RAPID, SPONTANEOUS, AND INTERACTS WITH HIGHER-LEVEL VISUAL PROCESSING

1. Introduction

In order to successfully navigate a perceptually chaotic world and share our understanding of it with others, we must not only extract the identity of people and objects, but also the roles that they play in events: Boy-hitting-girl is very different from girl-hitting-boy even though the event category (i.e. hitting) and actors involved are the same. In the former, the boy is the Agent (the actor) and the girl the Patient (the one acted upon), while in the latter, their roles are reversed. The fundamental importance of such "thematic roles" has long been emphasized in linguistics: Theories of thematic roles were initially developed to account for the consistent semantic properties of grammatical arguments (e.g., Subjects and Objects) across linguistic descriptions of events (Croft, 2012; Dowty, 1991; Fillmore, 1968; Gruber, 1965; Kako, 2006; Levin & Rappaport-Hovav, 2005) but now they are also a component of some theories of conceptual representation (Jackendoff, 1990; Langacker, 1987; Talmy, 2000), development (Baillargeon et al., 2012; Leslie, 1995; Muentener & Carey, 2010; Yin & Csibra, 2015), and perception (Leslie & Keeble, 1987; Strickland, 2016) more generally.

1.1. Event role extraction

While there is ongoing debate within linguistics about the precise number and nature of thematic roles in language, here we are interested in whether the mind, independently from explicit language production and comprehension tasks, rapidly and spontaneously extracts role information from perceptual input. Our work takes inspiration from a wealth of previous literature that has demonstrated rapid and bottom-up encoding of semantic content from visual scenes. These studies have revealed that categories of both objects (Biederman et al., 1988, 1982; Thorpe et al., 1996) and places (Oliva & Torralba, 2001; Potter, 1976) can be recognized from brief displays (sometimes as little as 13 ms); that the computation itself is rapid – occurring within 100-200 ms (VanRullen & Thorpe, 2001); and that the computation is relatively automatic (Greene & Fei-Fei, 2014).

In previous work we have shown that, just as with object and place categories, event category and event role information is in principle available in a bottom-up fashion from very brief displays (Hafri, Papafragou, & Trueswell, 2013; see also Dobel, Diesendruck, & Bölte, 2007; Glanemann, Zwitserlood, Bölte, & Dobel, 2016; Wilson, Papafragou, Bunger, & Trueswell, 2011). However, it is not yet known whether encoding of event information is rapid: all tasks in previous studies (to our knowledge) explicitly required participants to make a post-stimulus judgment about what was happening in the scene. Thus, the computation itself (although based on a briefly presented visual stimulus) could conceivably have continued for several seconds, up until response to the post-stimulus probe. Additionally, the computation might have occurred only because of the explicit demands of the task, rather than being spontaneous.

1.2. Spontaneity and generality of role encoding

Here, we define a spontaneous process as any process that is executed independently of an explicit goal. Such a process could be automatic, in the sense that it is mandatory given certain input characteristics (Fodor, 1983), but it could also be spontaneous but not automatic in the sense that, under some conditions and with some cognitive effort, the process could be prevented from being executed (Shiffrin & Schneider, 1977). In the present work, we test for spontaneity of event role encoding.

Given the particular importance of event roles to event understanding, the spontaneity of such a process would be beneficial as we engage the social world, since at any given moment we may be performing other perceptual tasks, e.g., identifying objects or spatial properties of the scene. It would also prove useful to the young language learner tasked with mapping utterances to the events that they refer to (Gleitman, 1990; Pinker, 1989).

In both of these situations (social and linguistic), the utility of role information arises from its relative generality, i.e., the identification of commonality between the actors engaged in different events, such as *kicking* and *pushing* (Dowty, 1991; Jackendoff, 1990; Pinker, 1989; Talmy, 2000; White, Reisinger, Rudinger, Rawlins, & Durme, 2017). However, research on action recognition using psychophysical and neuroscientific methods has largely focused on how the perceptual system differentiates between different action categories (e.g., *kicking, pushing, opening*) and generalizes within action

category (Hafri, Trueswell, & Epstein, 2017; Jastorff, Begliomini, Fabbri-Destro, Rizzolatti, & Orban, 2010b; Oosterhof, Tipper, & Downing, 2012a; Tucciarelli, Turella, Oosterhof, Weisz, & Lingnau, 2015; M. F. Wurm & Lingnau, 2015). This research has not yet addressed how we come to recognize the distinct roles that multiple actors play in visual scenes, or how (and whether) our perceptual system generalizes across the agents of different actions.

Investigating the perception of events in visual scenes provides an ideal avenue to test hypotheses about the generality of event roles. One hypothesis is that awareness of event-general properties of event roles (e.g., volition or cause) arise through explicit and deliberate observation of commonalities among event-specific roles (e.g., *kicker, kickee*) outside of the domain of perception (Tomasello, 2000). However, to the degree that we can find evidence that perception itself rapidly and spontaneously furnishes such event-general role information, the notion of event-specific roles as drivers of event understanding from scenes becomes less plausible. We hypothesize that in initial scene viewing, the perceptual system rapidly categorizes event participants into two broad categories – "Agent-like" and "Patient-like" (denoted Agent and Patient from here on for simplicity; Dowty, 1991; Strickland, 2016) – even if these assignments are later revised or refined in continued perceptual or cognitive processing of the event (see section 6.1 for elaboration on these issues).

1.3. The current study: an event role switch cost?

The goal of the current work is to establish the degree to which the visual system gives the observer event roles "for free", as part of routine observation of the world. We aim to show the following: (1) that the visual system encodes event roles spontaneously from visual input, even when attention is otherwise occupied (i.e. even when the observer is not explicitly asked to recognize events but rather is engaged in some orthogonal task); (2) that the computation of role itself is rapid; (3) that this encoding of event roles is at least partly event-general; and (4) that any evidence we find for encoding of event roles cannot be fully accounted for by simple visual correlates of event roles alone, such as posture.

To achieve this goal, we employed a "switch cost" paradigm (Oosterwijk et al., 2012; Pecher, Zeelenberg, & Barsalou, 2003; Spence, Nicholls, & Driver, 2001). In several

experiments, participants observed a continuous sequence of two-person scenes and had to rapidly identify the side of a target actor in each (Experiments 1a and 1b: male or female actor; Experiments 2 and 3: blue- or red-shirted actor). With our design, event role identities provide no meaningful information for the primary task of gender or color identification, so observers need not attend to such irrelevant information. Nevertheless, we hypothesized that when people attend to the target actor to plan a response, then if event roles are spontaneously encoded, they should "come along for the ride." Thus, we should be able to observe an influence of this role encoding on responses even though event roles are irrelevant to the primary task.

More specifically, we reasoned that if role assignment is spontaneously engaged, then when the role of the target actor switched from trial to trial, it would result in a cost, i.e., a relative lag in reaction time, even though subjects were tasked with identifying a property orthogonal to roles (here, gender or shirt color). If such a pattern were observed, it would provide compelling evidence that analysis of event structure from visual scenes is a rapid, spontaneous process that is engaged even when we are attending to other perceptual information. Furthermore, by using simple tasks based on visual information known to be rapidly available (including gender; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000), we expected that observers would respond quickly, allowing us to test the rapidity of extraction of event role information.

2. Experiment 1a

Participants observed a series of simple still images displaying an interaction between a male and a female, and were simply asked to say whether the male/female was on the left or right of the screen. We predicted that although the task fails to actively encourage role encoding (and may even discourage it), participants would nevertheless be slower on trials in which the event role of the target actor differed from his or her role in the previous trial, i.e., a "role switch cost".[2]

[2] We cannot differentiate between switch costs vs. repetition benefits (priming) because there is no baseline for comparison, but in keeping with the terminology in previous investigations using this paradigm (e.g., Pecher et al., 2003), we use the term switch costs. Whether the effects are a benefit or cost does not qualitatively change our conclusions.

2.1. Method

2.1.1. Participants

Twenty-four members of the University of Pennsylvania community participated and received either class credit or $10. Because we were collecting a large number of trials within-participant (see section 2.1.3 below), we predicted that this number of participants would be sufficient to observe the role switch cost, if it were to exist. All participants in this experiment and in the other experiments reported below gave informed consent, following procedures approved by the University's institutional review board.

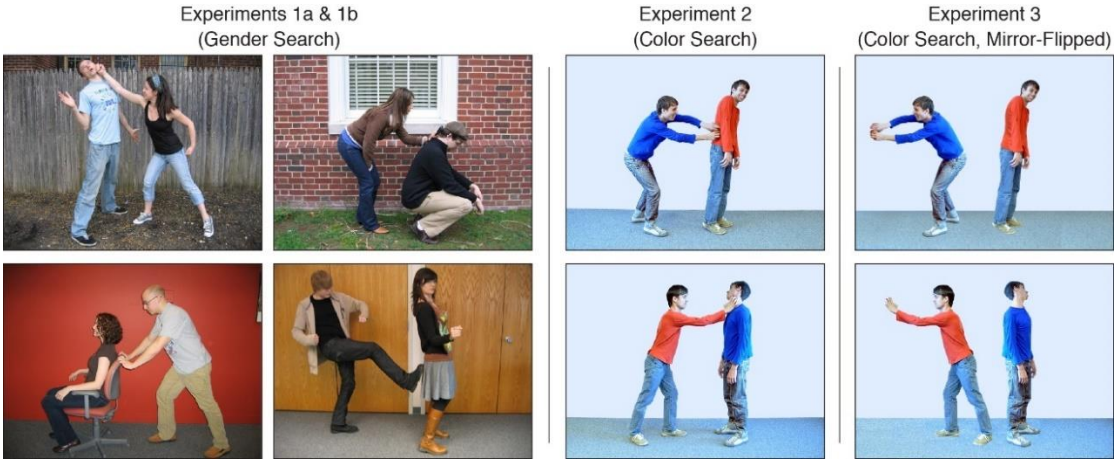


Figure 2.1
Example stimuli. All experiments featured 10 event categories. In Experiments 1a and 1b, these were depicted by several different pairs of actors, and Agent gender (male or female) and Agent side (left or right) were fully crossed within event category. In Experiments 2 and 3, events were depicted by a pair of identical twin actors. Agent shirt color (blue or red) and Agent side (left or right) were fully crossed within event category. In Experiment 3, the images from Experiment 2 were manipulated such that the two actors were always facing opposite one another; thus, their interactive relationship was almost entirely eliminated. See Appendix A for more example images.

2.1.2. Materials

The stimuli were 40 color photographic images depicting 10 two-participant event categories taken from a previous study that investigated extraction of event categories and roles from briefly displayed and masked images (Hafri et al., 2013). The event categories used were *brushing, chasing, feeding, filming, kicking, looking, punching, pushing, scratching, tapping*. These categories were chosen because they showed the highest agreement among subjects for role assignment from brief display (i.e., male as

Agent or Patient). All stimuli were normed for event category and role agreement in the previous study.

Six different male-female actor pairs appeared in the images, with each actor pair appearing in front of a different indoor or outdoor scene background. Each event category was associated with only one of the actor pairs (e.g., *brushing* and *chasing* was always performed by Pair 1, *feeding* by Pair 2, etc.). For each event category, the gender of the Agent (male or female) and the side of the Agent (left or right) were fully crossed, such that there were four stimuli for each event category. Each event depicted the actors in profile view. Example images appear in Figure 2.1, and examples for each event category appear in Appendix A.

For all experiments, images were 640 × 480 pixels and subtended 19° × 15° at approximately 54 cm distance from the screen. Stimuli were displayed on a 19" Dell 1908FP LCD monitor at a refresh rate of 60 Hz. Responses were collected using a PST E-Prime button box (mean latency 17.2 ms, *SD* 0.92 ms, measured in-lab). The experiment was run in Matlab using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

2.1.3. List design

Given that detecting switch costs depends on measuring the influence of one stimulus on another, we implemented "continuous carryover" sequences, which are similar to randomized block and Latin square designs, with the added benefit of controlling for first-order carryover effects, i.e. each stimulus precedes and follows every other stimulus (Aguirre, 2007; Nonyane & Theobald, 2007). This design resulted in 1601 trials split among 40 blocks. Unique lists were generated for every participant. An additional reason we used this list design was that it naturally provided a large number of trials with which to precisely measure effects of all factors manipulated in the experiment, across both subjects and items. This was important: given that participants were actively required to attend to stimulus features orthogonal to the property of interest (event roles), there was potential for the role switch cost to be quite subtle.

To maximize our chances of finding a switch cost if it were to exist, a small number of catch trials (Event Test trials) were randomly dispersed among the standard image trials. On these catch trials, participants were given a 2AFC test on what action just appeared in the previous trial (e.g., *kicking* or *pushing*). One label was correct, and the other was a foil randomly selected from the set of nine other categories. There were 58 catch trials in

total, with 1 to 3 per 40-trial block.

2.1.4. Procedure

Subjects were instructed that as each image appeared, they would have to press one of two buttons (left or right) to indicate, as quickly and accurately as possible, which side of the screen that the target actor appeared on (left button for left, right button for right). For half of the subjects, the target was the male actor, and for the other half, the female actor (i.e. male or female search was between-subject, counterbalanced across participants). There were 40 blocks of trials, each of which was a continuous sequence of all 40 image trials and the interspersed catch trials, followed by a quick break before the next block. The purpose of the catch trials was to focus participants' attention on the events they were observing without explicitly testing them on event roles (see section 2.1.3 above). Subjects were told that they would be intermittently tested on what action just appeared in the previous trial.

Figure 2.2 illustrates the trial and block sequence. Each trial consisted of the following: A "Ready?" screen for 350 ms, a central fixation crosshair for 250 ms, a blank screen for 150 ms, and the test image, which remained on the screen until the subject responded. Catch trials involved a similar sequence, but in place of the test image was a slide with the text "What action did you just see?" and two event category labels on either side of the screen (e.g., "biting" and "pushing"). Subjects selected their answer by pressing either the left or right button. Image trials timed out if no response was given within 2000 ms, and catch trials within 3500 ms. Twelve practice image trials and two catch trials preceded the main experimental sequence. Practice image trials depicted joint or symmetrical actions (e.g., *crying, shaking hands*). Average duration of the experiment was 41 min (which was similar across all additional experiments reported below).

[Manuscript continues with figure on next page]

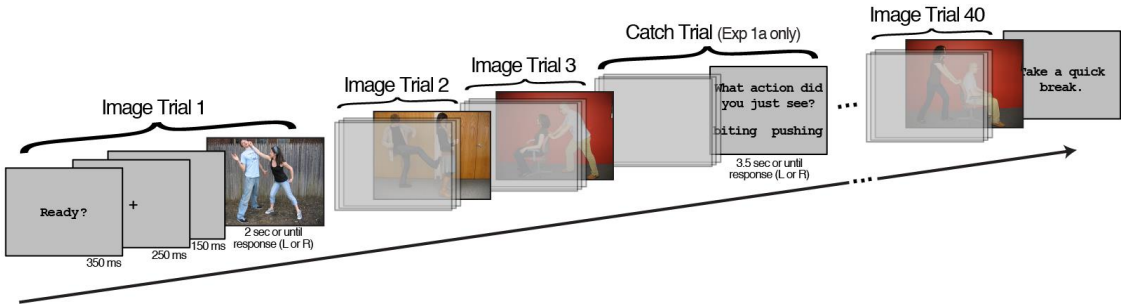


Figure 2.2

Block structure for all experiments. On each image trial, subjects pressed a button to indicate the position of the target actor as fast as possible (left or right). In Experiments 1a and 1b, the target actor was the male or female (between-subject). In Experiments 2 and 3, the target actor was the blue- or red-shirted actor (between-subject). Only Experiment 1a contained catch trials, which asked about the action that appeared in the previous trial.

2.1.5. Data analysis

Trial exclusion criteria were decided in advance of analysis and were the following: trials with an incorrect response and those following an incorrect trial, RTs faster than 200 ms, timeouts, trials after breaks, and trials after catch trials. An additional 63 trials in total across all subjects were also excluded due to an error in list creation. For the remaining data, trials with RTs 2.5 standard deviations above or below each subject's mean were also excluded, following accepted data trimming procedures (e.g., Balota, Aschenbrenner, & Yap, 2013). A mean of 17% (*SD* 4.0%) of trials in total were excluded per subject, which meant there were an average of 269 trials included per subject. Average accuracy was 95.6% (*SD* 2.2%), and average RT on image trials for the included data was 383 ms (*SD* 34 ms).

Individual trial reaction times from the primary task (i.e., judging gender side) were analyzed with linear mixed effects modeling using the lme4 R package (Bates et al., 2016), with centered (sum-coded) predictors. The analyses used the maximal subject and item random effects structure that converged for all tested models (Barr, Levy, Scheepers, & Tily, 2013).[3] RTs were first transformed into inverse RTs (-1000/RT) to

[3] When more complex random effects structures failed to converge, we successively dropped random slope terms with the smallest variance, until the model converged (Barr et al., 2013). The random effects structures used for each experiment and cross-experiment comparison were the following (in R model syntax):
Experiment 1a: (1+Actors+Side|subjNum)+(1+propertyAgent*sideAgent|event)
Experiment 1b: (1+Actors*Role+Actors*Side|subjNum)+(1+propertyAgent*sideAgent| eventCategory)
Comparison of Experiments 1a and 1b: (1+Role|subjNum)+(1+propertyAgent*sideAgent| eventCategory)

improve normality for model fitting. Additionally, all models included nuisance regressors for trial number and preceding trial inverse RT to account for general temporal dependencies (Baayen & Milin, 2010).

The following factors were included in models: Actors (repeated vs. switched), i.e., whether the actor pair was the same or different from the previous trial; Side (repeated vs. switched), i.e., whether the side of the target actor (e.g., male) was the same or different as the previous trial; and the effect of primary interest, Role (repeated vs. switched), i.e., whether the role of the target actor was the same or different (e.g., whether the male remained the Agent or switched to being Patient). Significance of these factors was tested by comparing likelihood-ratio values for nested models that included main effects and interactions of factors to models without them.[4]

2.2. Results

2.2.1. Role switch cost

An event role switch cost was observed. As shown in Table 2.1, participants were on average 6 ms slower when the role of the target character changed from one trial to the next. This effect, though quite small, was significant: The best-fitting mixed effects model included a main effect of Role (the role switch cost) and main effects and interactions of Actors and Side. The fit of this model was significantly better than a model without the main effect of Role, $\chi^2(1) = 52.9$, $p < .001$. Models with additional interaction terms were not a significantly better fit, either for Actors × Role ($\chi^2(1) = 1.71$, $p = .19$), or Side × Role ($\chi^2(1) = 0.09$, $p = .76$). See Table 2.1 for a summary of the effects from the best-fitting model.

2.2.2. Absolute vs. relative magnitude of role switch cost

Experiment 2: (1+Role*Side|subjNum)+ (1+propertyAgent*sideAgent|eventCategory)
Experiment 3: (1+Role|subjNum)+(1+propertyAgent*sideAgent|event)
Comparison of Experiments 2 and 3: (1+Role|subjNum)+(1+propertyAgent*sideAgent| eventCategory)
Abbreviations (consistent for all experiments): subjNum = subject identity; propertyAgent = Agent gender (Male or Female, Experiments 1a and 1b only), or Agent Color (Blue or Red, Experiments 2 and 3 only); sideAgent = Agent side (Left or Right); eventCategory = event category (e.g., *kicking*).

[4] Here and in Experiment 1b, repeated event always entailed repeated actors, due to the nature of the stimuli employed (see section 2.1.2). However, similar results were obtained with Event as a factor instead of Actors. Likewise, since actors and scene backgrounds co-varied, Actor switch entails a Background switch (and vice-versa), but for simplicity, we will refer to this factor as same/different Actors.

Before continuing, we believe that the empirical robustness and theoretical import of the role switch cost must be separated from the absolute size of the effect observed. Although the absolute magnitude of the role switch cost was small (about 6 ms), the *standardized* effect sizes were quite large: Cohen's *d* of 1.07 and 2.24, for subjects and items, respectively (see Figure 2.3). As another indication of its robustness, 21/24 participants and all 10 event categories showed a numerical difference in line with the role switch cost. And while it may be surprising that such a small effect would be statistically significant, each observer provided on average 1329 data points, resulting in very stable performance estimates per subject and per item (e.g., note the tight 95% confidence intervals across subjects in Table 2.1). Furthermore, it is within the same order of magnitude of previously observed switch costs, relative to mean RTs for task: for example, Pecher et al. (2003) obtained a cost of 29 ms relative to mean RTs of 1139 ms (a ratio of 2.5%), and Oosterwijk et al. (2012) obtained a cost of 22 ms relative to mean RTs of 1683 (a ratio of 1.3%), compared with our 6 ms vs. 383 ms mean RTs (a ratio of 1.6%). Similar arguments hold regarding the absolute vs. relative magnitude of the role switch cost observed in the other experiments reported in this manuscript, and we return to this issue in section 6.6.

2.2.3. Other observed effects

Besides the effect of primary interest (event roles), the best fitting model revealed several additional effects. First, people were slower when Actors switched. This is not surprising: when actor pair switched, it likely took longer to ascertain which character was the male or female. There was also an interaction of Side × Actors: On trials where the actor pair was different, participants were faster when the target side switched. Though speculative, it may be that with a significant visual change such as a switch in the actors, subjects may have expected a side switch, resulting in a switch benefit, or faster RTs. Whatever the reason for these additional effects, the role switch cost was invariant to these other factors (Side and Actors).

2.2.4. Event catch task

Average RT on catch trials was 1177 ms (*SD* 215 ms), and accuracy on catch trials was significantly above chance across participants (mean = 85%, *SD* = 10%, $t(23) = 40.0$, $p < .001$, $d = 3.37$, 95% CI = [81%, 89%]). This indicates that participants were monitoring

the events in the images sufficiently to distinguish which of two event categories they observed in the previous trial.

One important question is whether event category extraction is related to event role extraction. Although in our previous work we found that role recognition was not significantly correlated with event category extraction on an item-by-item basis (Hafri et al., 2013), we can also address this in the current study, in two ways. First, if there is a relationship between event category and event role extraction, we might find that the magnitude of the role switch cost is correlated on a subject-by-subject basis with performance on catch trials (event identification). However, we found no significant correlation between individual participants' role switch cost magnitude (based on the mean inverse RT difference between repeated and switch role trials for each subject), and either their overall accuracy on catch trials, r = -0.11, $t(22)$ = -0.52, p = .61, or their mean inverse RT on catch trials (accurate trials only), r = 0.00, $t(22)$ = -0.01, p = .99.

Another way to investigate the relationship between event category and event role extraction is by asking whether catch trial (event identification) performance would be worse when the catch trial probe is about an image in which event role switched. To assess this, we split catch trials by whether the previous trial was a Repeated or Switched Role image trial (an average of 27.8 trials in each condition per subject, range 20-36). We ran multilevel models to predict performance (either accuracy or inverse RT) on catch trials across subjects. Specifically, we tested whether adding a main effect of Previous Role (Repeated vs. Switched) to the models would improve model fit, over a null model without the Previous Role main effect (both models included a random intercept and random slope for Previous Role for each subject). However, adding Previous Role did not significantly improve fit either for catch trial accuracy (logistic regression, $\chi^2(1) = 0.64$, p = .42) or for catch trial inverse RT ($\chi^2(1) = 3.09$, p = .08); and even though improvement for the inverse RT model was marginal, it went in the opposite direction of the prediction, i.e. faster RTs on catch trials when the previous trial role *switched*.

Although these tests are post-hoc and we should interpret the null results with caution, they at least imply that at the individual subject or trial level, event category identification is robust to changes in role information. Nevertheless, a more definitive test of the relationship between event role and category extraction would require further

experimentation.

Table 2.1

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	380 (14.2)	386 (14.8)	6 (2.00)	7.27*
Actors	371 (12.8)	385 (14.9)	14 (3.68)	3.99*
Side	390 (16.0)	377 (13.7)	-13 (6.26)	-3.95*
Side, Repeated Actors	371 (12.9)	371 (13.5)	0 (6.39)	0.47
Side, Switched Actors	393 (16.5)	378 (13.9)	-15 (6.55)	-3.95*

Mean RTs across subjects for Experiment 1a, separately for all factors that were significant in model fitting (significant interaction terms split by each factor level). 95% confidence intervals in parentheses. * $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 2.2.1 for details on model comparisons.

[Manuscript continues with figure on next page]

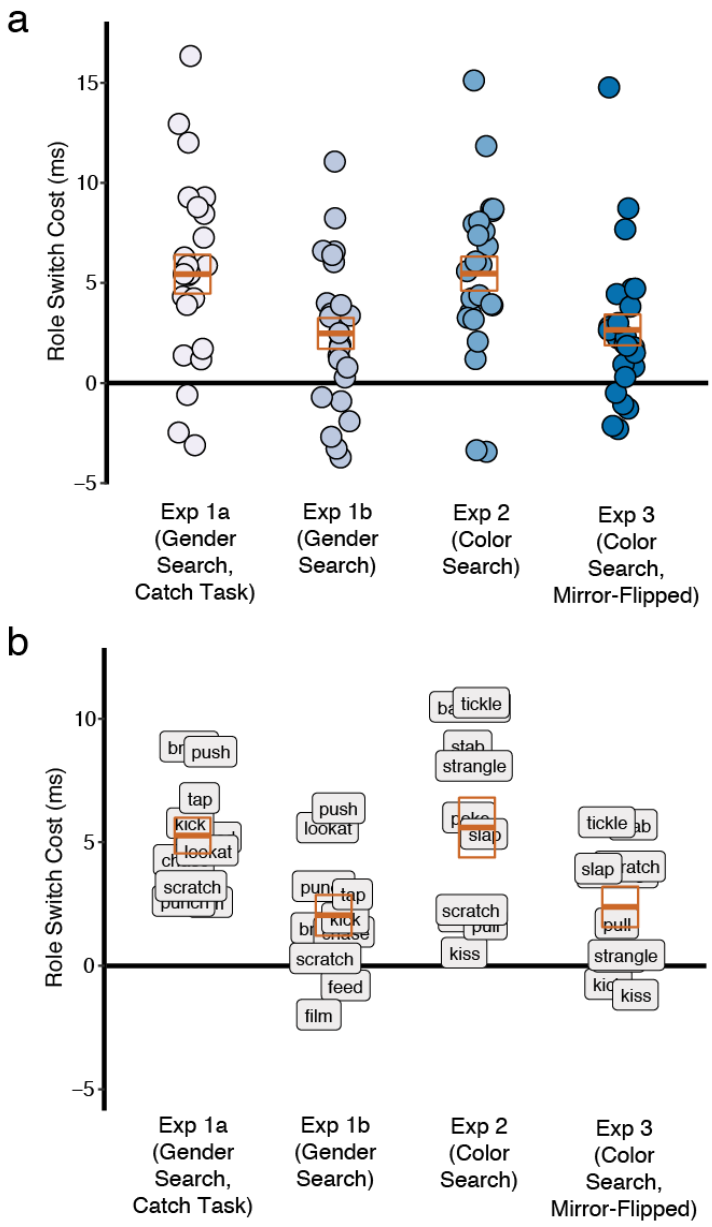


Figure 2.3
Individual (a) subject and (b) item (event category) means for the role switch cost, across all experiments. These plots show the consistency of the role switch cost for both subjects and items: the majority of means are above zero in each case. Orange boxes indicate the mean and standard error across subjects and items, for each experiment.

2.3. Discussion

Although a role switch cost was observed in Experiment 1a, the Event Test catch trials may have inadvertently focused attention on event roles. Experiment 1b was

identical to the previous experiment, except that there was no catch task and no mention of events or actions. If this effect is really a result of the default in visual perception of scenes, then we expected to observe it even under these conditions.[5]

3. Experiment 1b

The Event Test catch trials in Experiment 1a may have inadvertently focused attention on event roles. The current experiment was identical to Experiment 1a, except that there was no catch task and no mention of events or actions.

3.1. Methods

3.1.1. Participants

An additional 24 members from the University of Pennsylvania community participated and received class credit. Given the stability of the role switch effect in Experiment 1a, we believed this number to be sufficient.

3.1.2. Materials and procedure

All materials, apparatus, and procedure were identical to Experiment 1a, except that no catch (Event Test) trials were included, and instructions were modified to omit mention of the catch task or actions and events.

3.1.3. Data analysis

Data coding, trial exclusion criteria, and analysis were the same as in Experiment 1a. An additional 216 trials across all subjects were excluded due to an error in list creation. A mean of 13% (*SD* 4.9%) of trials (214 on average) per subject were excluded, average accuracy was 96.0% (*SD* 2.6%), and average RT for the included data was 387 ms (*SD* 48 ms). Individual trial RTs were analyzed using linear mixed effects modeling.

3.2. Results

As in Experiment 1a, a role switch cost was observed. In Table 2.2, we see that participants were on average 3 ms slower when the role of the target character changed from one trial to the next. This effect was once again robust: 17/24 subjects and 7/10

[5] Preliminary analyses of Experiments 1a and 1b originally appeared in conference proceedings (Hafri, Trueswell, & Strickland, 2016).

event categories went in the direction of the effect (Cohen's *d* of 0.55 and 0.58, respectively; see Figure 2.3). And although small, it was significant: The best-fitting mixed effects model included main effects and interactions of Role and Actors, as well as a main effect of Side and the interaction of Side × Actors. The fit of the model was significantly better than the same model without the additional interaction of Role × Actors, $\chi^2(1) = 4.89$, $p = .03$, and significantly better than a model that did not include Role at all, $\chi^2(2) = 15.5$, $p < .001$. A model with an additional interaction of Role × Side was not a significantly better fit, $\chi^2(1) = .004$, $p = .95$.

Interestingly, the role switch cost was greater when the actor pair repeated than when it did not, although importantly, the role switch cost was significant even when the actor pair differed. And as in Experiment 1a, on trials where the actor pair was different, participants were *slower* when the side repeated. See Table 2.2 for details.

Table 2.2

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	385 (19.9)	388 (20.3)	3 (1.59)	2.62*
Actors	371 (16.8)	390 (20.8)	19 (5.55)	2.08*
Side	394 (19.5)	380 (21.2)	-14 (7.55)	-5.09*
Role, Repeated Actors	368 (16.8)	374 (17.2)	6 (5.65)	3.60*
Role, Switched Actors	388 (20.5)	391 (21.0)	3 (1.84)	2.62*
Side, Repeated Actors	368 (14.0)	374 (20.2)	6 (11.6)	0.92
Side, Switched Actors	398 (20.6)	382 (21.5)	-16 (7.26)	-5.09*

Mean RTs across subjects for Experiment 1b, separately for all factors that were significant in model fitting (significant interaction terms split by each factor level). 95% confidence intervals in parentheses. * $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 3.2 for details on model comparisons.

3.2.1. Comparison of Experiments 1a and 1b

Not surprisingly, more participants and items showed the numerical difference in Experiment 1a (with the catch task) than in Experiment 1b (without the catch task; 21/24 vs. 17/24 participants, and 10/10 vs. 7/10 items, respectively; see Figure 2.3). To formally compare the two experiments, we ran new mixed effects models with the data from both experiments combined, starting with a base model whose main effects and interactions were identical to the best-fitting model of Experiment 1b. The best-fitting model in this combined analysis had main effects of Actors, Side, Role, and Experiment,

and interactions of Actors × Side, Role × Actors, Role × Experiment, and Actors × Experiment. The fit of the model was significantly better than a model without the additional interaction of Role × Experiment, $\chi^2(1) = 3.88$, $p = .05$. The greater role switch cost for repeated actors vs. switched actors observed in Experiment 1b appears to be consistent across both Experiments 1 and 1b: adding the triple interaction of Role × Actors × Experiment to the best-fitting model in the current analysis did not significantly improve the fit, $\chi^2(1) = 0.74$, $p = .39$. This analysis confirms that the role switch cost was indeed greater in Experiment 1a than in Experiment 1b.

Additionally, items drove the role switch cost consistently across experiments: the role switch costs for individual image stimuli were correlated across experiment, $r = 0.37$, $t(38) = 2.43$, $p = .02$. This correlation further attests to the stability of the measures of central tendency (i.e., subject and item means) – likely due to the large number of observations per image.

4. Experiment 2

In this experiment, we tested the generalizability of the role switch cost. We ran the same paradigm of Experiment 1b, with two changes: (1) we used new event categories and stimuli, in which events were depicted by a pair of red- and blue-shirted identical twin actors; and (2) the main task was to identify the side of the blue or red-shirted actor. As in Experiment 1b, there was no catch task and no mention of events or actions. If spontaneous role assignment is really the default in scene perception, then we expected to observe the role switch cost even with these changes.

4.1. Methods

4.1.1. Participants

An additional 24 members from the University of Pennsylvania community participated and received class credit. Given the stability of the role switch effect in Experiments 1a and 1b, we believed this number to be sufficient. Data from an additional three participants were excluded: one for a high number of sub-200 ms RTs (145 trials), one for non-completion, and one for falling asleep.

4.1.2. Materials

Stimuli were 40 color photographic images depicting 10 two-participant event

categories, taken from a previous study (Hafri et al., 2013): *bandaging, kicking, kissing, poking, pulling, scratching, slapping, stabbing, strangling, tickling*. All categories except *kicking* and *scratching* differed from those used in Experiments 1a and 1b, providing a test of the generalizability of the role switch cost to new event categories. All stimuli were normed for event category and role agreement in the previous study, and showed high agreement for event role extraction from brief display. Events were depicted by a single pair of identical-twin actors (male, age 29) who dressed the same except for a difference in shirt color (blue vs. red). As in Experiments 1a and 1b, for each event category, the shirt color of the Agent (blue or red) and the side of the Agent (left or right) were fully crossed, such that there were four stimuli for each category. Example images appear in Figure 2.1, and examples for each event category appear in Appendix A.

4.1.3. Procedure

Apparatus, list design, and procedure were identical to Experiment 1b, except that that the words "male" and "female" were replaced by "blue" and "red" in the instructions.[6] Task (blue or red search) was between-subject, counterbalanced across participants. Sixteen practice trials using additional stimuli (e.g., *brushing*) preceded the main experiment. To make the color task comparable in difficulty to the gender task, images were desaturated using Photoshop software to a level of 3% (a level of saturation which made the color task more difficult but kept the actors distinguishable).

4.1.4. Data analysis

Data coding procedures and trial exclusion criteria were the same as in Experiments 1a and 1b. A mean of 14% (*SD* 4.9%) of trials (237 on average) per subject were excluded based on the previous exclusion criteria. Average accuracy was 96.2% (*SD* 2.7%), and average RT for the included data was 347 ms (*SD* 38 ms). Individual trial RTs were analyzed using linear mixed effects modeling with Event (repeated vs. switched), Side (repeated vs. switched), and Role (repeated vs. switched) as factors.

4.2. Results

[6] For Experiments 2 and 3, one extra repetition for each image stimulus (e.g., *kick-blue-left* → *kick-blue-left*) was included in case we found a need to examine exact image repetitions, but these were discarded a priori before analyses.

As in Experiments 1a and 1b, a role switch cost was observed. In Table 2.3, we see that participants were on average 6 ms slower when the role of the target character changed from one trial to the next. This effect was again robust: 22/24 subjects and all 10 items went in the direction of the effect (Cohen's *d* of 1.42 and 1.40, respectively; see Figure 2.3). And although small, this effect was significant: The best-fitting mixed effects model included main effects of Role, Side, and Event, and interactions of Role × Side and Event × Side. The fit of the model was significantly better than the same model without the additional interaction of Role × Side, $\chi^2(1) = 4.03$, $p = .04$; and significantly better than a model that did not include Role at all, $\chi^2(2) = 31.9$, $p < .001$. Additionally, a model that also included an interaction of Role × Event was not a significantly better fit, $\chi^2(1) = 1.22$, $p = .27$.

Interestingly, the role switch cost interacted with repeated side, such that the role switch cost was greater when the side repeated than when it did not; importantly, however, the role switch cost remained even when the side was different. Like the additional effects observed in Experiments 1a and 1b, participants were faster when the side repeated, but only when the event category repeated. See Table 2.3 for a summary of the effects from the best-fitting model.

To summarize, a role switch cost was once again observed, even when the stimuli, event categories, and task were different. In fact, several participants reported that they explicitly tried to ignore the action as part of their strategy in performing the color task, but nevertheless, nearly all participants demonstrated the role switch cost. The results from this experiment suggest that the role switch cost is a general and robust phenomenon.

[Manuscript continues with figure on next page]

Table 2.3

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	344 (16.2)	350 (16.4)	6 (1.75)	4.69*
Event	350 (16.6)	347 (16.2)	-3 (2.10)	-2.76*
Side	346 (16.1)	348 (16.9)	2 (6.00)	0.08
Role, Repeated Side	343 (16.0)	349 (16.3)	6 (2.41)	7.04*
Role, Switched Side	346 (16.9)	351 (17.0)	5 (1.89)	4.69*
Side, Repeated Event	344 (15.9)	353 (17.5)	9 (7.38)	2.60*
Side, Switched Event	346 (16.1)	348 (16.9)	2 (6.05)	0.08

Mean RTs across subjects for Experiment 2, separately for all factors that were significant in model fitting (significant interaction terms split by each factor level). 95% confidence intervals in parentheses. * $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 4.2 for details on model comparisons.

4.2.1. Does Agent saliency drive the role switch cost?

Although the findings thus far provide evidence for a role switch cost, such a cost could be driven solely by a switch from Agent to Patient or vice-versa (i.e. it could be asymmetrical). Indeed, Agent primacy and saliency effects have been observed in both the linguistics and vision literature: Agents tend to precede Patients in linguistic utterances (Dryer, 2013; Goldin-Meadow, So, Ozyürek, & Mylander, 2008), and in continuous event perception, Agents attract attention, likely because they initiate movement before Patients (Abrams & Christ, 2003; Mayrhofer & Waldmann, 2014; Verfaillie & Daems, 1996) or because active body postures direct spatial attention (Freyd, 1983; Gervais, Reed, Beall, & Roberts, 2010; Shirai & Imura, 2016).

If Agent saliency is driving the role switch cost, we should observe two additional effects in our data across experiments: (1) different average RTs on trials in which the target was the Agent (Agent judgment trials) as compared to trials in which the target was the Patient (Patient judgment trials); and (2) an asymmetry in the role switch cost, such that the cost for an Agent→Patient switch should be different from the cost for a Patient→Agent switch. Note that the directionality of the predictions (i.e. whether Agent trials should be faster or slower) depends on different theories about the interaction

between event perception and the building of event structure. Under the view that Agents attract attention because of their active posture or movement initiation (e.g., Gervais et al., 2010; Verfaillie & Daems, 1996), one would predict faster RTs to Agent trials relative to Patient trials, since the primary task of participants was to locate the target actor. Under the view that observing Agents triggers the building of an event structure (Cohn & Paczynski, 2013; Cohn, Paczynski, & Kutas, 2017), attending to Agents (i.e. Agent judgment trials) might result in an additional cost due to initiation of event structure building, and therefore slower RTs. The crucial point here is that for Agent saliency (whether faster or slower) to explain the role switch cost, an asymmetry should also be observed between Agent→Patient and Patient→Agent switch trials, not only a difference between Agent and Patient judgment trials.

To formally test for these effects, we ran new mixed effects model comparisons in which we added Trial Judgment (Agent or Patient judgment trials) to the best-fitting models described in the above Results sections, separately for each experiment. Differences between Agent and Patient trials would manifest as a main effect of Trial Judgment, and an asymmetry in the role switch cost would manifest as an interaction of Role × Trial Judgment.

For Experiments 1a and 1b, adding a main effect of Trial Judgment or a Role × Trial Judgment interaction to the previously best-fitting models did not offer a significant improvement (all p's > .11). For Experiment 2, adding a main effect of Trial Judgment did significantly improve the fit over the previous best-fitting model ($\chi^2(1) = 55.5$, $p < .001$): Agent trial RTs were *slower* than Patient trial RTs (349 ms vs. 345 ms in subject means; see Table 2.4). The slower Agent RTs in Experiment 2 are in line with the hypothesis that Agents may trigger the process of "event building" (Cohn & Paczynski, 2013; Cohn et al., 2017). However, adding an additional interaction of Role × Trial Judgment to this model was not a significant improvement ($p > .66$). Given that differences between Agent and Patient trials was not consistent across experiments and that an asymmetry was not observed, these analyses suggest that Agent saliency cannot account for the role switch cost observed in the previous experiments.

Table 2.4

Experiment	Reaction time (ms)		Agent trial advantage (ms)	*t* value for parameter in best-fitting model
	Agent trials	Patient trials		
Exp 1a (Gender Search, Catch Task)	383 (15.3)	383 (13.7)	0 (2.59)	0.58
Exp 1b (Gender Search)	387 (20.6)	387 (19.6)	0 (2.27)	1.58
Exp 2 (Color Search)	349 (16.3)	345 (16.2)	-4 (1.77)	-7.45*
Exp 3 (Color Search, Mirror-Flipped)	353 (24.7)	362 (23.1)	9 (2.91)	15.9*

Mean RTs across subjects for each experiment, split by Trial Judgment type (Agent and Patient judgment trials, i.e. whether the target actor was the Agent or the Patient on each trial). 95% confidence intervals in parentheses. * $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 4.2.1 for details on model comparisons.

4.3. Discussion

Experiment 2 replicates and extends the findings from Experiments 1a and 1b by showing that role switch costs can be observed in explicit tasks other than those involving judgments about gender. Thus, these effects seem to be quite general.

5. Experiment 3

In a final experiment, we probed the level of representation at which the role switch cost operates, testing two non-mutually exclusive possibilities. The first possibility, and the one of central theoretical interest to our investigation of event roles, is that the cost operates at the *relational level*: Agent and Patient roles are fundamentally relational (an Agent acts on a Patient), so perhaps it is the roles that scene entities take in an *interactive relationship* that results in the role switch cost. An alternative possibility, however, is that the role switch cost operates at the *pose level*: active body postures are probabilistically associated with Agents and not Patients (Hafri et al., 2013), so perhaps observed switch costs merely reflect salient changes in posture of the actors. Note that effects of posture, if they contribute to the switch cost, should have an equal effect whether the actors in the scene are interacting or not.

To test these two possibilities (*pose and relational levels*), we ran the same paradigm of Experiment 2, with one change: images were edited such that the actors always faced opposite directions ("mirror-flipped"). With this manipulation, the actors' poses were preserved but their interaction was substantially reduced or eliminated (see also

Glanemann et al., 2016). Thus, any switch costs observed in the current experiment (with non-interactive actors) can only be attributed to switches at the *pose level*.

The image manipulation in the current experiment will allow us to assess the specific contribution that two levels (*pose and relational levels*) make to the switch costs observed in our previous experiments. If the previously observed role switch costs were due only to informational conflict at the *relational level*, we should observe a complete elimination of the switch cost here, since any interaction between actors is now minimally present. If the switch costs were due only to the *pose level*, then there should be no consequence of the image manipulation: all and only the previous role effects should obtain. However, if the role switch cost in previous experiments was due to conflict at both levels (*relational* and *pose*), the switch cost should still obtain here, but its magnitude should be significantly lower than that of the switch cost in this experiment's closest counterpart (Experiment 2).

5.1. Methods

5.1.1. Participants

An additional 24 members from the University of Pennsylvania community participated and received class credit. Given the stability of the role switch effect across Experiments 1a, 1b, and 2, we believed this number to be sufficient. Data from an additional four participants were excluded: two for not completing the experiment and two for low accuracy (<86%). This accuracy threshold was based on performance of participants in the previous experiments (all >89%), although inclusion of these excluded participants did not qualitatively change the results.

5.1.2. Materials and procedure

Stimuli from Experiment 2 were edited in Photoshop such that actors always faced away from one another. This was achieved by flipping each actor (or both) horizontally about his own center axis. Since actors sometimes partially occluded one another (e.g., in *slapping*, the Agent's hand and Patient's face), this procedure occasionally resulted in missing body or face parts in the images. The missing regions were replaced with parts from other images using various Photoshop tools. This was successful: no subject noticed the image manipulation even when questioned during debriefing. Example images

appear in Figure 2.1, and examples for each event category appear in Appendix A. Apparatus and procedure were identical to Experiment 2.

5.1.3. Data analysis

Data coding procedures and trial exclusion criteria were the same as in Experiments 1a, 1b, and 2. A mean of 12% (SD 3.2%) of trials (190 on average) per subject were excluded based on the previous exclusion criteria. Average accuracy was 97.7% (SD 1.6%), and average RT for the included data was 358 ms (SD 56 ms). Main analysis procedures were the same as in Experiment 2. Although in principle the actors were no longer Agents and Patients due to the mirror-flip manipulation, we coded Role (repeated vs. switched) based on each actor's corresponding role in the unedited stimuli.

5.2. Results

A role switch cost was once again observed. In Table 2.5, we see that participants were on average 3 ms slower when the role of the target character changed from one trial to the next. This effect was robust here as well: 20/24 subjects and 8/10 items went in the direction of the effect (Cohen's *d* of 0.86 and 0.97, respectively; see Figure 2.3). And although small, it was significant: The best-fitting mixed effects model included main effects of Role and Side. The fit of the model was significantly better than the same model that did not include Role at all, $\chi^2(1) = 13.8$, $p < .001$. Additionally, a model that also included an interaction of Role × Side was not a significantly better fit, $\chi^2(1) = 0.10$, $p = .75$, nor was a model that also included a main effect of Event, $\chi^2(1) = 0.01$, $p = .92$. As in the previous experiments, participants were slower when side repeated. See Table 2.5 for details.

Table 2.5

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	356 (23.5)	359 (24.2)	3 (1.59)	3.86*
Side	363 (26.1)	353 (22.0)	-10 (6.81)	-15.8*

Mean RTs across subjects for Experiment 3, separately for all factors that were significant in model fitting. 95% confidence intervals in parentheses. * $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 5.2 for details on model comparisons.

5.2.1. Comparison of Experiments 2 and 3

Given that Experiments 2 and 3 are a minimal pair, they present an ideal opportunity for additional assessment of the contributions of the *pose* and *relational levels* to the role switch cost. Because of the mirror-flip manipulation in the current experiment, the role switch cost here can only be attributed to the *pose level* (since the interaction between actors was minimal or non-existent), while in Experiment 2 it can be attributed to both *pose* and *relational* levels. Indeed, the size of the standardized effect in Experiment 3 was about two-thirds of that observed in Experiment 2 (see Tables 2.3 and 2.5). To formally compare the role switch cost across experiments, we ran new mixed effects models with the data from both experiments, with a base model whose random effects structure, main effects, and interactions were identical to the best-fitting model of Experiment 3. Adding a main effect of Experiment and interaction of Role × Experiment to the base model significantly improved the fit as compared to a model with only a main effect of Experiment, $\chi^2(1) = 10.6$, $p = .001$. This comparison yields credence to the idea that a combination of levels (*pose* and *relational*) led to the switch costs observed in Experiment 2.[7]

5.2.2. Does Agent saliency mediate the role switch cost in this experiment?

Here, unlike in previous experiments, there was a reliable Agent trial advantage: participants were on average 9 ms faster to respond on Agent judgment than Patient judgment trials. This was confirmed in mixed effects models: adding Trial Judgment (Agent vs. Patient judgment trial) as a factor to the best-fitting model from above significantly improved the fit, $\chi^2(1) = 252$, $p < .001$. Furthermore, this Agent advantage was greater than in any other experiment (independent samples *t* tests over subjects: all *t*'s > 6.40, *p*'s < .001; paired samples [Experiment 2] and independent samples [Experiments 1a and 1b] *t* tests over items: all *t*'s > 3.31, *p*'s < .01; see Table 2.4 for the

[7] In the mirror-flip manipulation, it could be argued that the interactive nature of the actors is not completely eliminated; for example, a kicker facing away from a would-be kickee may appear instead to be marching away from the other actor – a kind of social interaction. If this is the case, the reduced effect here could be due to a reduction (but not full elimination) of the interaction between actors, rather than a combination of the relational and pose level information. However, based on responses to questions during debriefing, the majority of participants considered the actors non-interacting. Thus, although the role switch cost in this experiment should perhaps be called a "posture switch cost", we use the term "role switch cost" for consistency with the previous experiments.

magnitude of Agent advantage in each experiment). As discussed in section 4.2.1, for Agent saliency to account for the results here, we would also expect an asymmetry in the role switch cost, i.e. a differential cost for Patient-switch than Agent-switch trials. However, this additional effect was not observed: adding an interaction of Role × Trial Judgment did not improve model fit over a model with only a main effect of Trial Judgment, $\chi^2(1) = 2.02$, $p = .16$. Thus, we can conclude that Agent saliency (or more properly here, "active posture" saliency) did not mediate the role switch cost in the current experiment.

The contrast in directionality of the Agent saliency effects between Experiments 2 and 3 is further evidence that these stimuli were analyzed at different levels (*pose* vs. *relational*) by the participants in each experiment. In Experiment 2, Agent trials were slower than Patient trials, consistent with the hypothesis of Agents triggering event-building in visually analyzed event scenes due to Cohn et al. (2013; 2017). In the current experiment (Experiment 3), we speculate that a different process may be at work: the actors were analyzed at the postural level, with no event building initiated (given that actors in the scene were not interacting with one another). The robust effect of Trial Judgment (faster Agent judgment, or "active posture" trials) in this experiment is consistent with previous work that argues that active postures independently guide attention in scenes (Freyd, 1983; Gervais et al., 2010), even for infants (Shirai & Imura, 2016).

5.3. Discussion

We again observed a reliable role switch cost, but this differed substantially from our previous experiments. First, the effect size here was roughly two-thirds that of Experiment 2. Second, unlike in previous experiments, an Agent (active posture) advantage also obtained. Thus, the *pose level* alone (i.e., active and passive posture differences associated with certain roles) cannot account for the entirety of the role effects across studies. Instead, the role switch cost observed in previous experiments was likely operating at both the *pose* and *relational levels*.

Given the differences observed between Experiments 2 and 3, we propose that the perceptual system may be differentially attuned to interacting and non-interacting individuals. On the one hand, the perceptual system is likely tuned to active postures

generally, in line with evidence that active body postures direct spatial attention (Freyd, 1983; Gervais et al., 2010; Shirai & Imura, 2016). But for interactive events (Experiments 1a, 1b, and 2), we hypothesize that attention naturally spreads to both actors (the Agent and Patient). Indeed, recent work has shown a facilitatory effect on recognition of two-person interactions (relative to non-interacting dyads) akin to the well-known face-inversion effect, such that inversion effects are found for stimuli in which two people are facing each other but not when they are facing away (Papeo, Stein, & Soto-Faraco, 2017). Although our experiment was not explicitly designed to test for a general attentional advantage for interacting vs. non-interacting actors, we did find some evidence that such an advantage may exist. RTs were approximately 11 ms lower in Experiment 2 (in which actors were in interactive relationships, mean RT 347 ms) than in Experiment 3 (in which actors were mirror-flipped, i.e. not interacting, mean RT 358 ms). This was confirmed in a paired *t* test comparing RTs for individual image stimuli across the two experiments, collapsing over all cross-trial switch factors (e.g., the mean inverse RT for the image of *blue-kicking-red-from-the-left* in Experiment 2 compared to its mirror-flipped equivalent in Experiment 3), $t(39) = 11.5$, $p < .001$, $d = 1.83$.

Given that accuracy on the main task (color search) was actually numerically *higher* in Experiment 3 vs. Experiment 2 (97.7% vs. 96.2%, respectively), we do not believe the overall RT difference between the two experiments is due to general confusion on account of the mirror-flip manipulation; instead, the RT difference supports the hypothesis that there is an attentional advantage specific to interacting human figures, as if the perceptual system treats the interacting figures as an attentional unit.

5.3.1. Can the role switch cost be attributed to order effects or to the large number of trials used?

One general concern across experiments is that – although the large number of trials per subject (about 1600) resulted in robust estimates of central tendency – we might be capturing an effect that is due to the peculiarities of the task. This could surface as order effects: perhaps the role switch cost is due to effects of getting acquainted with the task (gender or color search), or perhaps it is an effect that emerges from overlearning the response to each stimulus, or to fatigue. We tested these possibilities directly, by adding additional interaction terms for Role (the switch cost) and either Trial Number (1 to approx. 1600) or Block Number (1 to 40) to the best-fitting model for each experiment.

Adding the Role × Trial Number interaction term did not improve any of the model fits, all $\chi^2(1) < 1.64$, p's > 0.20, nor did adding the Role × Block Number interaction term (with an additional main effect of Block Number), all $\chi^2(1) < 1.47$, $p > 0.23$. Thus, it seems unlikely that the role switch cost is driven by any peculiarities attributable to order effects, such as gradual accommodation to the task, overlearning, or fatigue.

Additionally, given that we obtained such a large number of observations per subject (about 1600), we wanted to ask whether we would have observed the role switch cost with fewer observations than were obtained in each experiment. To test this, we performed a power analysis that tested at which point in the experiment, if we had stopped collecting data, we would have found a significant role switch cost (at a standard significance level of $\alpha = .05$). Specifically, separately for each experiment, we performed identical mixed model comparisons to those reported in each experiment above, using the same best-fitting models (i.e., comparing the likelihood ratio values for models with and without Role as a factor). This was performed on data from each block, cumulatively (e.g., for Cumulative Block 1, this only included data from block 1; for Cumulative Block 2, data from both block 1 and 2; for Cumulative Block 3, data from blocks 1-3; etc., all the way up to block 40, which included data from the entire experiment). We simply asked at which block significance ($p < .05$) was reached and maintained for subsequent blocks in model comparisons. This is depicted in Figure 2.4. We find that for Experiments 1a and 2, as little as one-tenth of the data was sufficient to reach and maintain significance, and for Experiments 1b and 3, about half to two-thirds. Thus, we can be confident that in general, our estimate of the amount of data required was conservative, and we likely would have detected the role switch cost even with many fewer observations per subject and item.[8]

[8] We should note that the experiments reported in this manuscript were the first that we conducted using this switch cost paradigm, and the first (to our knowledge) to use this method in scene perception research in general. Therefore, given our initial uncertainty in how strong of an effect we should observe in such a paradigm, we used a large number of trials per subject to maximize our chances of observing an effect of event role if it were to exist. Since we found that only a subset of the trials was needed to detect the role switch cost in our experiments, we hope that the reported power analysis proves useful to other researchers interested in using a similar paradigm for asking questions about encoding of event information in visual scenes.

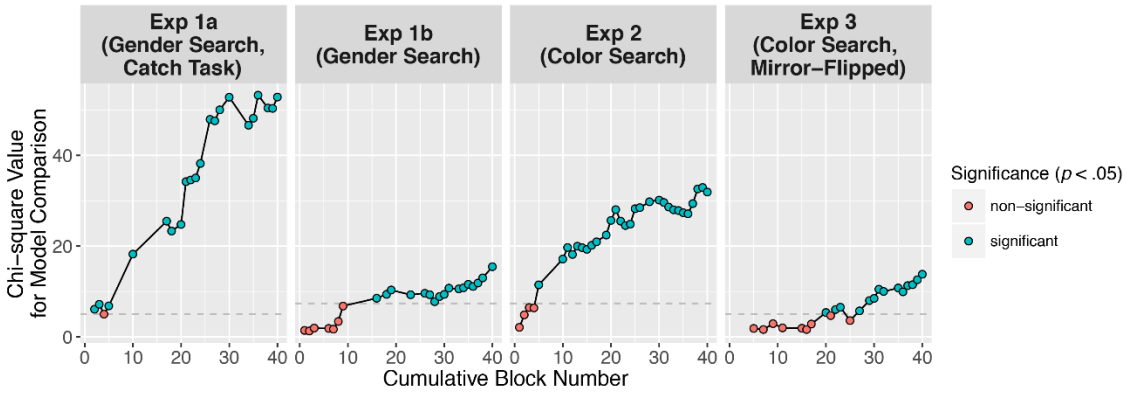


Figure 2.4

Analysis of the amount of data required to obtain a significant role switch cost effect in each experiment. Mixed effects model comparisons (for models with and without Role as a factor) that were identical to those reported for each experiment were calculated on data from each block, cumulatively (i.e., for *cumulative block number* on the x-axis, each block number also includes data from all previous blocks, e.g., the data point for block number 30 represents a statistic calculated using models with data from blocks 1-30). The dotted line in each plot indicates the chi-square value required for a level of significance of $p < .05$ for that experiment's model comparison. Blue points indicate significant chi-square values. (Points for some data subsets do not appear because models using those subsets did not converge). These plots indicate that fewer blocks of trials would have been sufficient for detecting the role switch cost in each experiment (in some cases, such as in Experiments 1a and 2, we would have detected the switch cost with as little as one-tenth of the data).

5.3.2. Can linguistic encoding of our stimuli explain the role cost?

Given that there is evidence of rapid interplay between event apprehension and utterance formulation (Gleitman, January, Nappa, & Trueswell, 2007), it is conceivably possible that linguistic encoding of the stimuli was happening, even within this short time frame (<400 ms). If the switch cost we observed is due to purely *grammatical* categories (Subject, Object), then our experiments cannot adjudicate the generality of event roles (i.e., Agent and Patient, or related cluster-concepts; Dowty, 1991; White et al., 2017). In other words, *kicker* and *tickler* may not be conceptually related, but when they are situated in utterances, the *kicker* and *tickler* become similar by virtue of their both being grammatical Subjects (the same reasoning applies to *kickee* and *ticklee*).

However, linguistic encoding is unlikely to explain the role switch costs observed in our experiments for several reasons. First, explicit linguistic encoding was rare: in post-experiment questioning, only nine subjects across all experiments reported linguistically encoding the stimuli at any point in terms of who did what to whom (2 in Experiment 1a, 5 in Experiment 1b, 2 in Experiment 2, and 0 in Experiment 3). Second, any linguistic

encoding that occurred appears to have had little influence on the role switch cost: the cost was not statistically different between participants that reported encoding the events linguistically and those that did not, for any experiment (all *p*'s > 0.20, unpaired *t*-tests). In fact, only two of the nine participants that reported linguistic encoding, both in Experiment 1b, appeared in the top 50th percentile of switch cost magnitude among the other participants in their experiment.

It is also unlikely that participants were linguistically encoding the events implicitly. If they were, then we might expect a grammatical Subject advantage: Subjects appear first in utterances in English (a Subject-Verb-Object language), so trials on which the target actor was the Agent (the grammatical Subject in canonical active-voice utterances) might show faster RTs than when the target actor was the Patient (the grammatical Object). However, this was not the case: Agent (Subject) trials were actually significantly *slower* than Patient (Object) trials in Experiment 2, and there was no significant Agent (Subject) advantage in Experiments 1a and 1b (see Table 2.4).

Taken together, these analyses suggest that – although some participants did report encoding the stimuli linguistically – it had little if any influence on the role switch effects observed in our studies. Future work could further probe the influence of language on performance in a task such as ours by testing participants from different language groups, or those without access to fully formed natural language (e.g., deaf homesigners; Feldman et al., 1978; Zheng & Goldin-Meadow, 2002).

5.3.3. How general is the role switch cost over transitions between particular event categories?

In previous analyses, we found some evidence that the role switch cost is at least partly event-general (i.e. not tied to the specific preceding event category): in Experiments 1a and 1b, the role switch cost still held when Actor Pair (and therefore Event Category in that stimulus set) switched (see Tables 2.1 and 2.2 and section 3.2.1); and in Experiment 3, there was not a significant interaction of the role switch cost with repeated/switched event category. However, it still could be the case that the cost is dependent on which particular event categories precede others (i.e. that the role switch cost is driven by a small subset of preceding event categories). For example, in the extreme, it could be that the role switch cost for each event category is obtained only when preceded by the category *kicking*.

To address this, we simply calculated the average role switch cost (using inverse RTs) across subjects for each event category to every other event category, collapsing over Agent side. This yielded a 10 × 10 matrix of values for each experiment, where each cell of a matrix represents the average role switch cost for a transition from one particular event category to another, illustrated in Figure 2.5*A* (using raw RTs). We then tested whether these event-to-event role switch costs were significantly above zero for each experiment. Indeed as illustrated in Figure 2.5*B*, this was the case (all $t(99) > 2.39$, $p < 0.02$), even when excluding transitions between the same event categories, i.e. the diagonals of the matrices (all $t(89) > 2.03$, $p < 0.05$).[9] These analyses suggest that, at least for the event category exemplars used in our experiments, there is some commonality across the roles of the participants in different event categories that is driving the role cost. Implications for event role representations more broadly appear in section 6.1.

[Manuscript continues with figure on next page]

[9] The same analyses can be conducted using mixed effects models, testing whether the effect of Repeated Role no longer significantly improves model fit once event-to-event transitions are taken into account (operationalized here as separate random intercepts for Previous Event and the Previous Event × Current Event interaction, with random slopes for Repeated Role for each random intercept). These analyses support the same conclusion as the *t*-test analyses in the main text, namely that the role switch cost is not driven by a small set of event category transitions.

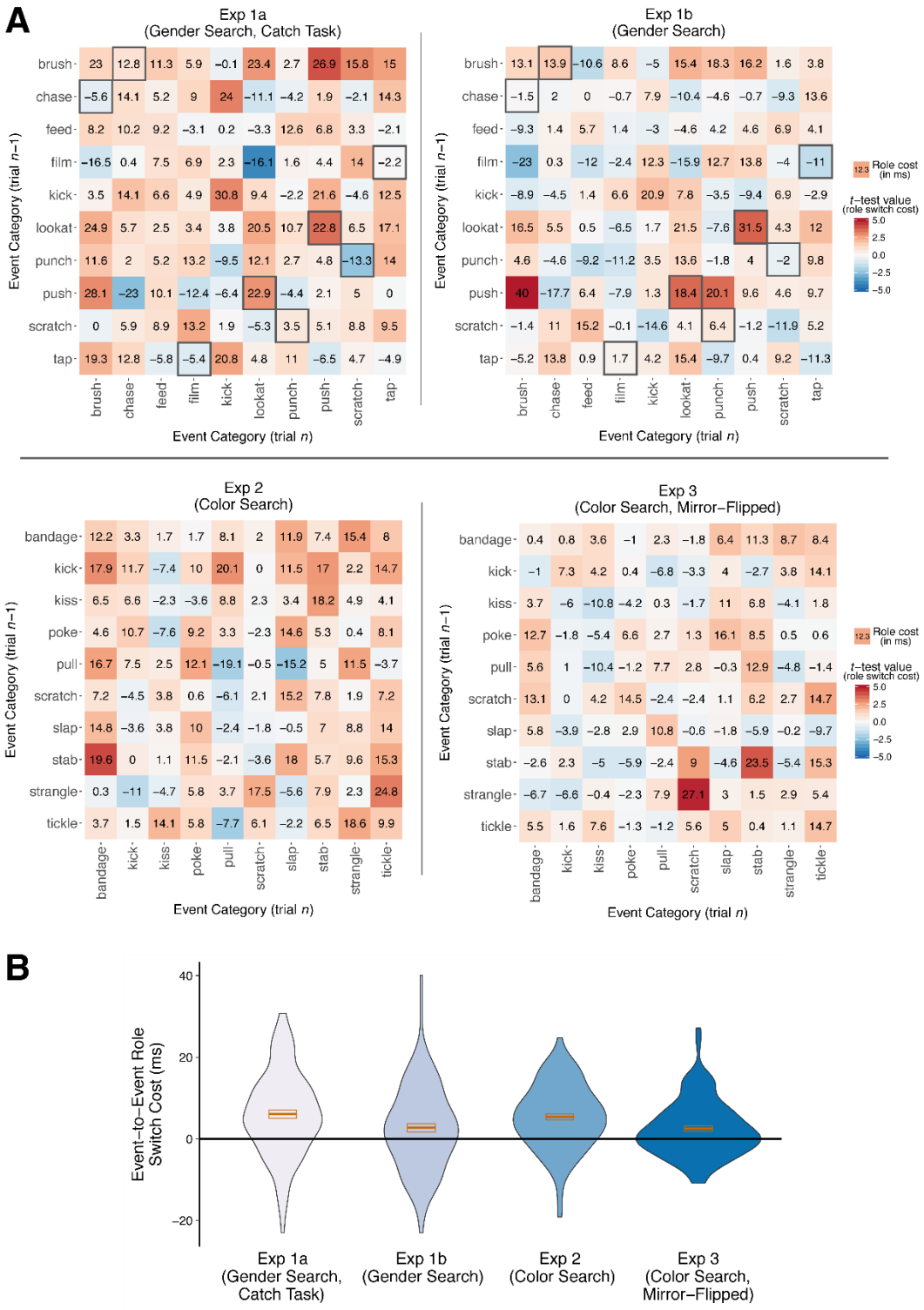


Figure 2.5

(a) Mean role switch cost over subjects (in milliseconds) calculated between each Event Category and every other Event Category, collapsing across Actor Side, separately for each experiment. Color shading indicates

t-test values for the switch cost across subjects ($|t(23)| > 2.07$ is significant at $p < .05$ uncorrected), with red indicating a role switch cost, and blue a role switch benefit. Gray boxes around cells in Experiment 1a and 1b matrices indicate transitions between different Event Categories that feature the same Actors (see section 2.1.2), which was found in analyses to result in higher switch costs (see section 3.2.1); this is not indicated in Experiments 2 and 3 since there was always only one set of actors. Note that diagonals in each matrix represent the switch cost for the same Event Category, so always reflected the same set of actors, in all experiments. (b) Violin plots of all cells from the four matrices in (a). Violin plot outlines indicate the kernel probability density, i.e. the width of each plot indicates the proportion of event-to-event transition values at each role cost magnitude. Orange boxes indicate the mean and standard error across transition values, for each experiment. Analyses showed that the role switch cost was not driven by a small subset of event-to-event transitions: as can be seen, the majority of values were above zero.

6. General Discussion

Our experiments demonstrate that the structure of an event, i.e. who acted on whom, is spontaneously encoded in visual processing, even when attention is directed toward other visual features (here, gender or color). This process manifested as a role switch cost, i.e., a relative lag in reaction time when the role of the target actor switched from trial to trial. The effect was robust across stimuli, event categories, and task (Experiments 1a and 1b: gender search; Experiment 2: color search). In Experiment 3, we determined that the role switch cost observed in the previous experiments cannot be fully explained by body posture differences associated with Agents and Patients. Furthermore, we found that the cost was not driven by a subset of the possible transitions from one event category to another, suggesting that the role information computed is quite general. Taken together, our experiments demonstrate (for the first time, to our knowledge) both the rapidity and generality of the event role computation itself.

6.1. Implications for event role representations

Although we have shown that assignment of Agent and Patient to entities in visual scenes is rapid and spontaneous, it may be that in continued processing, this coarse role assignment can be reversed or refined, in at least three ways. The first is additional visual input, in the form of successive fixations: for example, upon further observation, perhaps one recognizes that the initially identified Patient is holding a weapon, making him an Agent (a *shooter*); or that an Agent is holding an object to transfer, making the Patient a Recipient. Indeed, a recent gist-extraction study of event scenes revealed that observers need substantially longer viewing times to identify the coherence of spatially local event

properties such as the category of instrument objects vs. global event properties such as posture/orientation (Glanemann et al., 2016). The study of Glanemann et al.(2016) highlights the advantage afforded by initial commitment to a coarse role assignment: it can help guide scene fixations in a targeted manner (see also Castelhano & Henderson, 2007).

A second way that role assignment can be reversed or refined is via flexible event construal: Despite how an event plays out in the world, people can construe it in an innumerable number of ways (sometimes for comedic effect: "Yeah, I'm fine. I snapped my chin down onto some guy's fist and hit another one in the knee with my nose"; Ross, 1972). We speculate that in general, flexibility in event construal reflects a top-down, cognitive re-interpretation of an initial commitment provided rapidly by the visual system.

Finally, the context in which an event occurs likely allows for later assignment of more event-specific roles like *helper* or *hinderer* that incorporate this contextual information. Indeed, there is developmental evidence for both event-general and event-specific role distinctions: young infants readily distinguish Agents and Patients in social events like *helping* and *hindering*, but they also themselves prefer positively valenced Agents (i.e., helper; Hamlin et al., 2007; Kuhlmeier, Wynn, & Bloom, 2003).

Given that in our experiments, we found the role switch cost to be somewhat event-general, an important theoretical question is whether there are *systematic* differences in the role switch cost in terms of hypothesized properties of roles in different event categories. In particular, some theories of event roles hypothesize that certain components of events (e.g., contact, causation, and change of state or motion) are conceptual primitives, posited as such because they are relevant for grammar (Levin, 1993; Levin & Rappaport-Hovav, 2005; Pinker, 1989; Talmy, 2000) or because they are available early on in development (Strickland, 2016). Notably, these event components are similar to features proposed in cluster-concept notions of event roles (Dowty, 1991; Kako, 2006; White et al., 2017).

Although the consistency we observed in the role cost across events is broadly suggestive of generality (see Figure 2.5, and section 5.3.3), we do not believe we have a convincing way to address the precise characteristics of this generality with the current data, for the following reasons. First, the event categories we used did not independently

vary in theoretically relevant event components such as cause, contact, state-change, and motion. Second, we had essentially only one exemplar (i.e. one postural "tableau") per event category (see Figure 2.1 for examples). Thus, to address the generality and granularity of event roles extracted from visual scenes, future work will need to include many more event categories and to systematically manipulate hypothesized event components within event category.

Whatever theoretical distinctions end up accounting for the complexities of an observer's event conceptualization, we assert that there is a rapid and spontaneous assignment of Agent-like and Patient-like roles to interactive event participants, possibly before more refined role distinctions (e.g., Recipient) or social contingencies (as in the *helping/hindering* case) have been computed, and in some cases before event-specific role identification occurs (e.g., *kicker, kickee*).

Consequently, now that we have established the robustness and generality of the basic phenomenon of spontaneous role extraction with Agent-like and Patient-like event participants, there is a large set of theoretically interesting questions about how the visual system parses the roles in events with different numbers of participants and different relationships among them. For example, in single-participant events where the participant undergoes a change of state or location (e.g., *melting*, *falling*), is the participant assigned a Patient-like rather than Agent-like status? In a joint interaction such as *dancing*, are participants may be assigned similar roles (e.g. both Agents) rather than Agent and Patient? What is the role status of participants in complex events such as transfer events (e.g., *giving*, *serving*)?

6.2. Implications for the relationship between perceptual and linguistic encoding of event roles

The early stages of event perception as examined in the current studies have the potential to inform theories of argument selection in linguistic descriptions of events (i.e., whether event participants belong in sentential subject, object, or oblique positions). Our general theoretical viewpoint consists of the following notions: (1) in early perceptual processing, scene entities are categorized as Agent-like and Patient-like, often before the event category itself is determined; and as such, (2) initial role categorization is not dictated primarily by the event category itself (along with the

corresponding verb-specific roles such as Stimulus Experiencer, and Instrument), but rather by the perceptual particulars of the scene, i.e. the particular *token* of the event category. Our studies provide support for these notions: we found role switch costs even across exemplars of event categories that would not be considered in the literature to be canonical Agent-Patient relationships: events with a mediating instrument (*stab*, *film*, and *bandage*); events without caused motion or state-change (*look at*, *call after*, and *film*); and an event of transfer (*feed*), where the Patient might more traditionally be considered a Recipient.[10] Our viewpoint provides a possible perceptual explanation for at least two issues in linguistic argument selection: (1) the optionality and argument status of some event participants, such as Instruments; and (2) the cross-linguistic variability in grammatical status of certain event roles, such as Stimulus and Experiencer.

First, let us consider the optionality and argument status of event participants. It is debated whether instruments should be considered arguments of verbs: to describe a *stabbing* event, for example, one may say *John stabbed the man* or *John stabbed the man with a knife*. Rissman and colleagues (2015) account for these inconsistencies at the level of event construal: argumenthood depends on construal of a *particular token* of an event as indicated by a verb and its sentential context, rather than an absolutist notion of arguments that depends solely on the verb itself. Our work provides a perceptual complement to this notion: we argue that early available perceptual cues to role assignment have a strong influence on initial event construal. Hence, the degree of perceptual salience of objects involved in a particular token of an event should partially determine the degree to which an argument of a verbally encoded event will be optional, or should be considered an argument at all in the case of Instruments (see also Brown & Dell, 1987, on the pragmatics of inclusion of event participants in discourse).

The rapid and spontaneous encoding of event participants as Agent-like and Patient-like might also account for the fact that linguistic argument selection for certain event categories is more consistent cross-linguistically than for others. For example, the Agent-

---

[10] Of course, the scene exemplars (the images used for *look at*, *feed*, etc.) were selected precisely because there *was* general agreement in our previous study (Hafri et al., 2013) about the roles of the scene participants (who was performing the action vs. being acted upon). However, the fact that we found the role cost even for these items suggests that it is in principle possible to find Agent and Patient-like role effects even for categories of events without canonical Agent-Patient relationships. This provides evidence that the category of event does not exert a strong influence on early role assignment.

and Patient-like status of the subject and object in a description of a *hitting* event is fairly straightforward. In contrast, the statuses of subject and object in a description of a *frightening* or *fearing* event are much less clear (e.g., *John frightens Mary* and *Mary fears John* can describe the same event; Dowty, 1991), with some hypothesizing thematic roles distinct from Agent and Patient for these event participants (i.e., Stimulus or Experiencer, dependent on which participant is seen as the implicit cause of the event; Hartshorne, 2014; Levin & Rappaport-Hovav, 2005). We hypothesize that from instance to instance of a given event category, the Agent- and Patient-like perceptual properties of the participants may on average be less variable (e.g., *hitting, kicking*) or more variable (e.g., *fearing, frightening, looking*). Thus, it is not surprising that event categories involving Stimulus/Experiencer-like roles (e.g., *fearing*) are the ones for which there is high cross-linguistic variability in terms of which participant must appear in subject position. Indeed, we have previously argued that the high degree of cross-linguistic correspondence between Agents/Patients and subjects/objects is probably not a coincidence, but rather reflects a fundamental relationship between "core" cognition and perception (Strickland, 2016).

This brings us to the question of the degree to which language dictates conceptual event role assignment. It has certainly been shown that the linguistic framing of an event may influence attention to and memory for certain event participants or event components (e.g., Fausey, Long, Inamori, & Boroditsky, 2010; Kline, Muentener, & Schulz, 2013; Papafragou, Hulbert, & Trueswell, 2008; J. C. Trueswell & Papafragou, 2010). Notice here, however, that these phenomena reflect how language production alters attention in scenes ("looking for speaking"), or how language comprehension affects event construal (serving as a marker of the scene entities considered relevant by the speaker). We predict that cross-linguistic differences should be minimal in the first moments of event perception, and only afterward might language-specific effects be observed, if at all (e.g., language-specific conventions in terms of assignment of Stimulus and Experiencer to certain grammatical positions). Such a prediction could be tested by running our experimental paradigm with speakers of different languages, or with populations with no exposure to fully formed natural language, e.g. deaf homesigners (Feldman et al., 1978; Zheng & Goldin-Meadow, 2002). A second prediction is that an observer's event construal will be more susceptible to linguistic modulation when the

ambiguity of the initial role information available in the scene is higher, such as with Stimulus-Experiencer events, e.g. *frightening*, where the relative Agent-like or Patient-like cues between event participants may not significantly differ. In other words, speakers certainly use specific verbs and frames in an event description to convey the importance of the various event participants to their event construal (e.g., *frighten* vs. *fear*), but an observer's construal depends heavily on the perceptual parameters of the interaction *in the first place*.

To summarize this section, we believe that our results help to address some puzzles in the linguistic encoding of events, such as the argument status of event roles like Instruments and the cross-linguistic variability in grammatical status of certain roles like Stimulus and Experiencer. We speculate that in the first moments of event perception, how Agent-like and Patient-like scene participants are, as well as their perceptual salience, matters more for event construal and subsequent linguistic encoding than the logical relationship between event participants (such as Stimulus/Experiencer) in the depicted event category.

6.3. Implications for high-level visual perception

Our work is consistent with a wealth of previous literature that has demonstrated rapid, bottom-up encoding of semantic content from visual scenes (Biederman et al., 1982; Castelhano & Henderson, 2007; Greene & Fei-Fei, 2014; Greene & Oliva, 2009; Potter, 1976; VanRullen & Thorpe, 2001). Crucially, we find that not only are the perceptual features that are correlated with event role (i.e., body posture) extracted by the visual system rapidly, but the computation of the abstract role information itself is rapid. Observers in our studies viewed the scenes for less than 400 ms (based on mean response times), so for us to have obtained the role switch cost, the computation of role information must have taken place within this time frame.

Our findings fit within a broader literature in visual perception which shows that spontaneous and possibly automatic perceptual processes are not limited to low-level properties (e.g., lines and edges), but also extend to "high-level" representations that include objects (Scholl, 2001), event types (Strickland & Scholl, 2015), causality (Kominsky et al., 2017; Rolfs et al., 2013), and animacy (van Buren et al., 2015). Like our event role results, these other processes often map neatly onto representations from the

literature on infant "core cognition" and potentially conflict or diverge from higher-level, explicit judgments (Cheries, Wynn, & Scholl, 2006; Spelke & Kinzler, 2007; see Strickland, 2016, for a discussion of the relationship between elements of "core" cognition and cross-linguistic grammatical patterns). Additionally, the differences in the role switch cost for interactive actors (Experiment 2) and non-interactive actors (Experiment 3) supports the hypothesis that another element of core cognition that is reflected in perception are the social interactions of others, including their roles (Spelke & Kinzler, 2007). This is in line with other recent work suggesting that the perceptual system treats interacting figures as an attentional unit (Papeo et al., 2017) and that there is a region in the human brain selective for observed social interactions (Isik, Koldewyn, Beeler, & Kanwisher, 2017).

An open question is the extent to which the role switch cost is specific to human interactions, or is a reflection of more general processing of the interactive relationships between scene entities, both animate and inanimate. That is, in event scenes that involve interactions with or among inanimate objects (e.g., a woman opening a door or a ball hitting a rock), are roles assigned using similar visual processes? Given our assertion that early in visual processing, scene entities are assigned coarse Agent-like and Patient-like roles, it follows that, if an inanimate object is salient enough in the visual representation, it should also be rapidly assigned an Agent-like or Patient-like role. However, there is evidence that visual processing of animate and inanimate entities is quite distinct, both in terms of differential attention (van Buren et al., 2015) and underlying cortical pathways (Connolly et al., 2012; Scholl & Gao, 2013). It will require further investigation to determine whether the visual system assigns roles similarly to animate and inanimate scene entities.

6.4. Implications for action and event perception

Researchers studying action perception and its neural substrates have tended to focus on single-actor actions (e.g., *walking*; Giese & Poggio, 2003; Lange & Lappe, 2006) or actor-object interactions (e.g., grasping, opening; Rizzolatti & Sinigaglia, 2010; M. F. Wurm & Lingnau, 2015 and many others). Our work suggests that to gain a complete picture of action perception and the neural substrates supporting it, researchers must also study the event structure of actions and interactions (Hafri et al.,

2017). Additionally, our results have implications for theories of event perception from ongoing activity, particularly Event Segmentation Theory (EST; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). EST holds that during continuous perception, people construct an "event model" that includes relevant causes, characters, goals, and objects (Zacks, Speer, & Reynolds, 2009). Importantly, EST implies that this process does not require conscious attention. Our results directly support this core implication: people rapidly and spontaneously encode the structure of observed events, even when attention is guided to other properties of observed scenes. Our results further suggest that event roles should be considered key components of event models themselves, an intuitive notion: if event roles change, then so does the currently observed event.

6.5. Spontaneity vs. automaticity of role encoding

In the introduction to this paper, we defined a spontaneous process as any process that is executed independently of an explicit goal. Such a process could be automatic, in the sense that it is mandatory given certain input characteristics, but it could also be spontaneous but not automatic in the sense that, under some conditions and with some cognitive effort, the process could be prevented from being executed. Our results at minimum demonstrate the spontaneity of role encoding. However, what can we say about the potential automaticity of event role encoding?

One criterion for automaticity is the notion of "ballistic" engagement, i.e. that given certain types of perceptual input, a particular process is necessarily engaged and runs to completion (e.g., an English speaker cannot help but process the sounds of English as such; Fodor, 1983). Additional criteria are due to Shiffrin and Schneider (1977), who studied target item search among distractor items: they assert that automatic processing is quick, is not hindered by capacity limitations of short-term memory, and requires only limited attention. One difficulty in assessing the degree of automaticity using these criteria is that there is not a straightforward mapping between Shiffrin and Schneider's definitions of target and distractor and our definitions in the present study. In Shiffrin and Schneider (1977), targets and distractors are different objects (e.g., letters and numbers) on screen. In contrast, in our paradigm, the "target" (gender/color information) and "distractor" (role information) are two levels of description of the *same entity* (the target actor). Thus, if attention to *different levels* of the same stimulus and to

different stimuli should be considered analogous under the Shiffrin and Schneider criteria, then our results are consistent with automaticity: even when attention is directed to one level of the target actor (gender/color), we find that subjects also encode the same entity at another level (role).

However, since gender and color in our stimulus set were not in direct conflict with role information, only orthogonal to it, answering whether role extraction is automatic rather than simply spontaneous requires further research. Notably, such a distinction between spontaneity and automaticity is relevant not only within the domain of the current study, but applies to many fields investigating processes that have the potential to be considered automatic (e.g., theory of mind; Leslie, 1994; Scholl & Leslie, 1999).

6.6. Practical vs. theoretical significance of the role switch cost

Before we close, we believe that a separation of the empirical robustness, practical consequences, and theoretical import of the role switch cost is warranted. The empirical evidence is clear. We have reported a highly replicable effect, with each experiment showing a consistently large standardized effect size (minimum Cohen's *d* 0.55), and with a majority of subjects and items showing the effect in all cases. We also demonstrated that the large number of observations per subject were not necessary to obtain the effect (see Figure 2.4 and section 5.3.1).

For practical purposes, we are not surprised at the small absolute magnitude of the effect (about 5 milliseconds), since our experiments were explicitly designed to dis-incentivize people from making role categorizations. Remarkably, even under these fairly extreme conditions, participants exhibited a trace of tracking event roles. Nevertheless, we would expect whatever mental mechanisms that produce the tiny absolute effect sizes here to matter much more in everyday situations where Agency and Patiency *are* task-relevant (e.g. for the purposes of producing language or judging the behavior of conspecifics).

We assert that the theoretical importance of the effect is not measured by its absolute size, but rather by the theoretical distinctions made over the course of the experimental investigation. Indeed, despite its size, the stimulus manipulation of Experiment 3 provided evidence that the role switch cost is attributable not only to differences at the *pose level* (i.e., switches in body posture), but also to a more abstract *relational level*

(i.e., switches in event roles).

6.7. Conclusions

To close, over the course of four experiments, we have provided empirical evidence that the human visual system is spontaneously engaged in extracting the structure of what is happening in the world – including the interactive relationships between people. The rapidity of the extraction and its generality over a wide range of events suggests that this information may have a strong influence on how we describe the world and understand what we observe more generally.

III. NEURAL REPRESENTATIONS OF OBSERVED ACTIONS GENERALIZE ACROSS STATIC AND DYNAMIC VISUAL INPUT

1. Introduction

The ability to recognize actions performed by others is crucial for guiding intelligent behavior. To perceive categories of actions, one must have representations that distinguish between them (e.g., *biting* is different from *pushing*) yet show invariance to different instantiations of the same action. Although previous work has described a network of regions involved in coding observed actions (the "Action Observation Network", or AON; Caspers, Zilles, Laird, & Eickhoff, 2010; Kilner, 2011; Rizzolatti & Sinigaglia, 2010; Urgesi, Candidi, & Avenanti, 2014), the extent to which these regions abstract across differences between action exemplars is not well understood.

Previous research has addressed the question of abstraction (i.e. invariance) in two ways. First, many neuroimaging and neuropsychological studies have explored generalization between observed and executed actions, in an effort to resolve a debate over motor system involvement in action understanding (Chong et al., 2008; Dinstein et al., 2008; Kilner et al., 2009; Oosterhof et al., 2012a, 2012b; Tarhan et al., 2015; Tucciarelli et al., 2015; for review, see Rizzolatti and Sinigaglia, 2010; Oosterhof et al., 2013; Caramazza et al., 2014). Second, other studies have examined invariance to different perceptual instantiations of observed actions (Kable & Chatterjee, 2006; Oosterhof et al., 2012a; Tucciarelli et al., 2015; C. E. Watson, Cardillo, Bromberger, & Chatterjee, 2014). In an especially direct test of such invariance, Wurm and Lingnau (2015) found that representations in several AON regions distinguished between *opening* and *closing* in a manner that generalized across different kinematic manipulations and acted-upon objects (i.e., across bottles and boxes; see also M. F. Wurm, Ariani, Greenlee, & Lingnau, 2015). These findings and others suggest that at least a subset of AON regions support abstract codes for actions that could conceivably facilitate perceptual recognition.

However, we posited that an action recognition system should display two additional kinds of perceptual generalization. First, it should support representations of action category that are invariant not only to the acted-on object or kinematics, but also to

other incidental perceptual features, such as the identities of entities involved, and location. Whether a girl pushes a boy or a boy pushes a button, and whether it takes place in a classroom or a playground, it is still *pushing*. Second, these representations should be elicited both by dynamic visual sequences, in which the entire action is observed, and static snapshots, from which the causal sequence must be inferred (Hafri et al., 2013). Several studies have found action-specific representations using static images (Ogawa & Inui, 2011; C. E. Watson et al., 2014), but crucially, none have demonstrated common representations across dynamic and static input. Beyond testing these invariances, we also wished to examine actions performed with a wide variety of effectors (e.g., foot, mouth), not just hand/arm actions that are commonly investigated in the literature (Jastorff, Begliomini, Fabbri-Destro, Rizzolatti, & Orban, 2010a; Kable & Chatterjee, 2006; C. E. Watson et al., 2014).

To these ends, we used multivoxel pattern analysis (MVPA) of fMRI data to identify regions supporting abstract action representations. We scanned subjects while they viewed eight action categories, in two visual formats (Figure 3.1): (1) controlled videos of two interacting actors; and (2) still photographs involving a variety of actors, objects, and scene contexts. We then attempted to decode action category by comparing multivoxel patterns across the formats, which should be possible in regions that support action category representations not tied to low-level visual features correlated with actions. To anticipate, we were able to decode action category across visual formats in bilateral inferior parietal lobule (IPL), bilateral occipitotemporal cortex (OTC), left premotor cortex, and left middle frontal gyrus (mFG). We then conducted further analyses in these regions to probe the stability of their representations across perceptual features and subjects. Finally, we tested for action decoding in independently localized functional OTC regions to determine their involvement in action representation (Kanwisher, 2010). Taken together, our results support the hypothesis that AON regions contain neural populations that can mediate action recognition regardless of the dynamicity of visual input, and the perceptual details of the observed action.

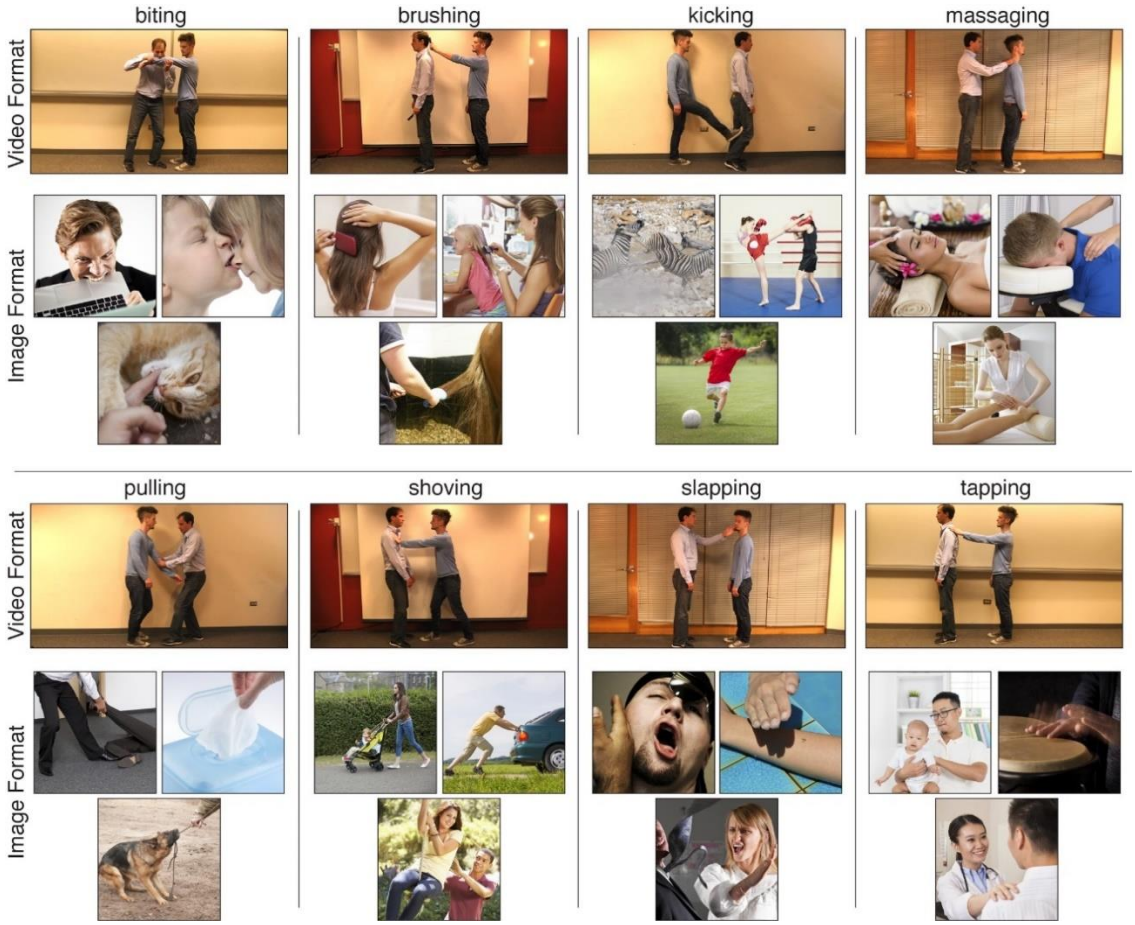


Figure 3.1

Examples of stimuli. Subjects viewed dynamic videos and still images of eight categories of interactions. For each action category, one still frame for the video format and three photographs for the image format are shown. In the video format, actor role (Agent/Patient), action direction (left/right), and scene background (four indoor backgrounds) were fully crossed within each action category. For example, in the *brushing* still frame depicted here, the blue-shirted actor is the Agent, the action direction is towards the left, and the background is the red wall, while in other *brushing* videos, this combination of factors was different (e.g., action direction towards the right instead of left). In the still image format, photographs from the internet were chosen to maximize variation in actors, objects, viewpoint, and scene context within each category. Image format examples shown here are photographs which we have license to publish and closely resemble actual stimuli used.

2. Materials & Methods

2.1. Participants

Fifteen healthy adults (8 female; mean age 22.1 ± *sd* 4.6 years; range 18-35 years) were recruited from the Penn community. All participants were healthy, had normal or corrected-to-normal vision, and provided written informed consent in compliance with

procedures approved by the University of Pennsylvania Institutional Review Board. All were right-handed, except one who was ambidextrous. All were native English speakers, and one was bilingual. Data from an additional participant was discarded before analysis for an inability to complete the entire experiment.

For selection of video stimuli, a group of 16 additional individuals (Penn undergraduates) participated in an online norming survey for psychology course credit. For selection of still image stimuli, 647 individuals on Amazon's Mechanical Turk (MTurk) participated in a separate online norming survey. All MTurk workers were located in the US and had 90% to 95% worker approval rate on previous tasks.

For the eyetracking control experiment, a group of 16 additional individuals (Penn undergraduates) participated for psychology course credit.

2.2. Stimuli

To identify neural representations of action categories that were invariant to incidental perceptual features, we scanned subjects while they viewed eight different categories of two-person interactions: *biting, brushing, kicking, massaging, pulling, shoving, slapping, tapping*.

The action categories were viewed in two formats: controlled video clips created in the lab, and visually varied photographic images taken from the internet. The use of videos allowed us to examine action representations elicited by dynamic stimuli, thus mimicking action perception in the natural world. This approach is the standard in previous literature investigating action recognition (e.g., E. D. Grossman & Blake, 2002; Vangeneugden, Peelen, Tadin, & Battelli, 2014; M. F. Wurm & Lingnau, 2015). The use of images allowed us to determine whether the same action representations were elicited even when actions are perceived from a static snapshot, which has been shown in previous behavioral studies to be sufficient for recognition even from brief displays (Hafri et al., 2013).

In addition, by using one format that was more visually controlled (the videos), and another that was more visually varied (the images), we decreased the possibility of potential confounding factors present in either format alone. The videos always contained the same set of actors and scene contexts, so the different body movement patterns were the only aspect of the stimuli that allowed categories to be discriminated

(apart from *brushing,* which contained a unique object). Although this had the merit that distinctions between categories within the videos could not be attributed to differences in actors or scene context, it had the disadvantage that category was inevitably confounded with lower-level motion properties that covaried with the actions. In the still images, on the other hand, distinctions between categories could not be attributed to low-level motion patterns; however, because the stimuli were less visually constrained, it remained possible that action category could have covaried with the presence of particular types of actors, objects, scene contexts, or even implied motion (Kourtzi & Kanwisher, 2000; Senior et al., 2000). By comparing patterns of fMRI responses to the videos with those to the still images when identifying category representations, we reduce these concerns, because the most likely confounds in one stimulus set are either absent or controlled for in the other.

2.2.1. Video stimuli

128 video clips (2.5 s each) were filmed, divided equally into eight action categories. A pair of male actors of similar height performed all interactions. Video clips were filmed in front of four different indoor backgrounds; one actor appeared as the Agent (i.e., the entity that performs an action on another entity) and the other as the Patient (i.e., the entity on which an action is performed); and the action was directed either towards the left or to the right. These three factors were crossed to make 16 video clips for each category: 4 backgrounds × 2 actor roles (A as Agent or B as Agent) × 2 action directions (leftward or rightward). For example, for *biting*, there were four video clips (with different backgrounds) of actor A on the left biting actor B on the right, four of A on the right biting B on the left, four of B on the left biting A on the right, and four of B on the right biting A on the left.

The two actors were centered in the video frame in full-body profile view and started each clip at rest with arms at their sides. For half of the action categories (*biting, pulling, shoving, slapping*), the actors faced one another, and for the other half (*brushing, kicking, massaging, tapping*), they both faced the same direction. For *brushing*, both actors always held a brush. Actors kept neutral faces throughout the duration of the videos. Example still frames for each action category appear in Figure 3.1.

To ensure that our videos could be easily interpreted as depicting the intended action categories, we obtained descriptions of our videos from a separate group of raters. These

participants viewed a random selection of 100 videos, one at a time, and provided a verbal label that in their opinion best described each action depicted (total 15 labels per video clip, *sd* 0.45, range 14-16). These verbal labels confirmed that our video clips depicted the intended action categories: all were described with the intended verbal label or close synonym >95% of the time. Synonyms included: for *biting: chomping, gnawing*; for *brushing: combing;* for *kicking:* none*;* for *massaging: rubbing;* for *pulling: yanking, tugging, grabbing, dragging;* for *shoving: pushing;* for *slapping: hitting, smacking;* and for *tapping: patting*.

2.2.2. Still image stimuli

For each action category, we used 16 still images (128 total), which were selected to maximize the within-category variety of actors, objects, and scene contexts (e.g., only one *biting* image included a person biting an apple). Stimuli included both animate and inanimate Patients (the entity on which an action is performed).

To create this stimulus set, an initial set of candidate stimuli were obtained from Google Images using search terms that included the target verbal label, close synonyms, and short phrases (e.g., *patting* or *patting on the back* for *tapping*, *combing* for *brushing*, *pushing* for *shoving*, *smacking in the face* for *slapping*). This search procedure yielded 809 images (87-118 images per category). To reduce this set, a group of MTurkers followed the same norming procedure as for the videos. Each viewed a random selection of 60 images, and provided a verbal label that best described each action depicted (total 16 labels per image, *sd* 1.6, range 11-20). Based on these labels, we eliminated images that did not have high name agreement with the target verbal label or close synonym. Synonyms included: for *biting: gnawing, tasting, eating*; for *brushing: combing;* for *kicking: kickboxing;* for *massaging: rubbing, back-rubbing;* for *pulling: yanking, tugging, grabbing, grasping, dragging;* for *shoving: pushing;* for *slapping: hitting, smacking, punching;* for *tapping: patting, poking, touching*. Name agreement was at least 87% for each *biting, brushing, kicking,* and *massaging* image. For the other categories (*pulling, shoving, slapping,* and *tapping*), the name agreement criterion was relaxed to a minimum of 75%, 75%, 64%, and 53%, respectively, in order to retain at least 16 images per category. This resulted in a set of 209 images (16-38 per category) with high name agreement.

We then calculated three measures to assess low-level visual similarity among the

remaining images, with the aim of choosing a final image set with maximal visual dissimilarity within each category. The first measure was the Gist model (Oliva & Torralba, 2001), which is a set of image descriptors that represent the energy at different spatial frequencies and scales. Image similarity was calculated as the correlation of descriptor magnitudes between each pair of images. The other two measures were the average HSV hue channel values for each image and average HSV saturation channel values for each image. With these three measures in hand, we ran 10,000 permutations in which we randomly selected a subset of 16 images per category and calculated, for each category, the average distance in Gist space between all 16 images, and the variance across images in the hue and saturation channels. Of these permutations, we selected the one with the greatest average within-category Gist distance and greatest with-category variance across images for hue and saturation. Across the final set of 128 images, we luminance matched the HSV value channel using the Matlab SHINE toolbox (Willenbockel et al., 2010), and converted the images back to RGB space. Examples for each action category appear in Figure 3.1.

2.3. MRI acquisition

Scanning was performed at the Center for Functional Imaging at the University of Pennsylvania on a 3T Siemens Prisma scanner equipped with a 64-channel head coil. High-resolution T1-weighted images for anatomical localization were acquired using a three-dimensional magnetization-prepared rapid acquisition gradient echo pulse sequence [repetition time (TR), 1620 ms; echo time (TE), 3.09 ms; inversion time, 950 ms; voxel size, 1 × 1 × 1 mm; matrix size, 192 × 256 × 160 mm]. T2*-weighted images sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired using a gradient echo echoplanar pulse sequence (TR, 3000 ms; TE, 30 ms; flip angle, 90°; voxel size, 3 × 3 × 3 mm; field of view, 192 mm; matrix size, 64 × 64 × 44). Visual stimuli were displayed at the rear bore face on an InVivo SensaVue Flat Panel Screen at 1920 × 1080 pixel resolution (diag = 80.0 cm, w × h = 69.7 × 39.2 cm). Participants viewed the stimuli through a mirror attached to the head coil. Images subtended a visual angle of ~11.7 × 11.7°, and videos subtended a visual angle of ~18.9 × 10.7°. Responses were collected using a fiber-optic button box.

2.4. Design and task

2.4.1. Main experiment

To determine BOLD response to action categories in different visual formats, participants were scanned with fMRI while viewing the still images and videos. Images and videos were presented in separate scan runs, with four runs per format (eight total), alternating in sets of two (e.g. image run 1, image run 2, video run 1, video run 2, image run 3, etc.). The format that appeared first was counterbalanced across participants. Within format, stimuli were identical within odd-numbered and within even-numbered runs (e.g., stimuli in video runs 1 and 3 were identical, stimuli in image runs 2 and 4 were identical, etc.). Thus, except for repetition trials (see next paragraph), each stimulus was shown a total of twice over the course of the experiment, in separate runs.

To ensure attention to the stimuli, participants were instructed to press a button whenever the stimulus on the current trial was exactly the same as the stimulus on the immediately-preceding trial (repetition trials). Importantly, this task could not be performed by attention to the action category alone. Trials occurred every three seconds in a rapid event-related design. Videos were displayed for 2500 ms, followed by a 500 ms inter-trial interval (ITI) with a white fixation cross centered on a gray background. Images were displayed for 1200 ms, followed by an 1800 ms ITI. Each scan run included 64 trials in which unique stimuli were shown (8 for each category), 8 repetition trials, and 12 null trials, in which participants viewed a blank screen with a fixation crosshair for 3 s (total duration 4 min 33 s per scan run). A unique pseudo-randomized sequence of stimuli was generated for each scan run using *optseq2* (*http://surfer.nmr.mgh.harvard.edu/optseq*; RRID:SCR_014363) with the following parameters: psdwin 0 to 21, nkeep 10000, focb 100, nsearch 200000. Five extra null trials were added at the end of each scan run to ensure we captured the hemodynamic response to the last stimulus in each run.

Video stimuli were divided such that odd video runs contained the videos with two of the four backgrounds and even video runs contained the videos with the remaining two backgrounds. Thus, each video run included two stimuli for each combination of action category, actor roles, and action direction (8 stimuli per action category in each video run). The combinations of background splits were cycled through for each subject (e.g.,

subject 1 had backgrounds 1 and 2 in odd runs and backgrounds 3 and 4 in even runs, subject 2 had backgrounds 1 and 3 in odd runs and backgrounds 2 and 4 in even runs, etc.). Image stimuli were assigned to odd and even runs with a unique split for each subject (8 images per category for the odd runs, and 8 per category for the even runs). Stimuli were displayed using a Macbook Pro laptop with Matlab version 2013b (MathWorks, Natick, MA; RRID:SCR_001622) and the Matlab Psychophysics Toolbox version 3.0.11 (Brainard, 1997; Pelli, 1997; RRID:SCR_002881).

2.4.2. Functional localizers

In order to determine the information content for action categories in functionally selective brain regions, all subjects completed three functional localizer scans in the middle of each scan session. The first localizer featured static image stimuli to identify regions responsive to different stimulus categories. This run consisted of 25 blocks (15 s long each; run duration 6 min 15 s) of static images of faces, objects, scrambled objects, bodies, and scenes. Blocks 1, 7, 13, 19, and 25 were null blocks with a blank gray screen and white crosshair. Images were presented for 800 ms followed each, with a 200 ms inter-stimulus interval. Subjects performed a one-back repetition detection task (two repetitions per block).

The second localizer featured dynamic stimuli to identify regions responsive to biological motion and basic motion (E. Grossman et al., 2000; E. D. Grossman & Blake, 2002; Vaina, Solomon, Chowdhury, Sinha, & Belliveau, 2001). This run consisted of 25 blocks (18 s long each; run duration 7 min 30 s) of intact point-light displays of single-person actions (e.g., *waving*, *jumping*), scrambled versions of these stimuli (in which motion patterns were preserved but starting position of points was randomized), and static point-light still frames randomly selected from the scrambled point-light videos. Blocks 1, 5, 9, 13, 17, 21, and 25 were null blocks with a blank gray screen and centered red fixation point. Stimuli were presented for 1500 ms each, with a 300 ms inter-stimulus interval. Subjects performed a one-back repetition detection task (one repetition per block). To create these stimuli, motion capture data were taken from the Carnegie Mellon Motion Capture Database (*http://mocap.cs.cmu.edu*) and animated using the Biomotion Toolbox (van Boxtel & Lu, 2013).

The third localizer featured linguistic stimuli to identify regions responsive to linguistic depictions of actions (design based on Bedny et al., 2008). This run consisted

of 20 blocks (18 s long each; run duration 6 min 36 s), in which verbs and nouns were presented visually to participants in separate alternating blocks. On each trial (2.5 s each), participants had to rate the similarity in meaning of two words presented sequentially (1 s each) by performing a button press indicating their response on a scale of 1 to 4. Words were a set of 50 motion verbs (e.g., *to stumble*, *to prance*) and 50 animal nouns (e.g., *the gorilla, the falcon*), approximately equated for similarity and difficulty (available in supplementary material in Bedny, Dravida, & Saxe, 2014). Words were randomly paired within block.

2.5. fMRI data analysis

2.5.1. Overview

Our primary goal was to identify representations of action categories that generalized across dynamic videos and static images. To identify brain regions supporting such representations, we implemented a whole-brain searchlight analysis of multivoxel responses to action categories shown in both movie and image format. Once these regions were identified, we performed several further analyses to determine the properties of the encoded action categories. First, we compared the cross-format searchlight results to results from within-format searchlight analyses to observe the degree of overlap of within- and cross-format decoding. Second, with the regions identified by the cross-format searchlights, we performed a more fine-grained analysis of the responses to the video stimuli, to test whether category representation elicited by videos generalized across actor role and direction of action. Third, we performed a representational similarity analysis within these regions to determine if their category spaces within each region were similar across subjects. Finally, to determine the relationship between functional selectivity and coding of action category, we tested for cross-format and within-format category decoding in a number of functional regions of interest defined based on univariate responses in localizer scans.

2.5.2. Data preprocessing

Functional images were corrected for differences in slice timing by resampling slices in time to match the first slice of each volume. Images were then realigned to the first volume of the scan, and subsequent analyses were performed in the subject's own space.

Motion correction was performed using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002). Data from the functional localizer scans were smoothed with a 5 mm full-width at half-maximum Gaussian filter; data from the main experimental runs were not smoothed.

2.5.3. Whole-brain analysis of cross- and within-format action category decoding

To search for action category information across the brain, we implemented a searchlight analysis (Kriegeskorte, Goebel, & Bandettini, 2006) of multivoxel patterns elicited by the 8 action categories in video and static image format. We centered a small spherical ROI (radius 5 mm, 19 voxels) around every voxel of the brain, separately for each participant, and then calculated a discrimination index within each sphere. This index was defined as the difference between the Pearson correlation across scan runs for patterns corresponding to the same action category in different formats (e.g., *kicking* in the video format with *kicking* in the image format) and the Pearson correlation across scan runs for patterns corresponding to different action categories in different formats (e.g., *kicking* in the video format with *brushing* in the image format). If this index is positive, this indicates that the searchlight sphere contains information about action category (e.g., Haxby, Gobbini, Furey, & Ishai, 2001). We then assigned the resulting value to the central voxel of the sphere.

To define the activity patterns, we used general linear models (GLMs), implemented in FSL (*http://fsl.fmrib.ox.ac.uk/fsl/fslwiki*; RRID:SCR_002823), to estimate the response of each voxel to each action category in each scan run. Each runwise GLM included one regressor for each action category (8 total), one regressor for repetition trials, regressors for six motion parameters, and nuisance regressors to exclude outlier volumes discovered using the Artifact Detection Toolbox (*http://www.nitrc.org/projects/artifact_detect*; RRID:SCR_005994). A high-pass filter (100 Hz) was used to remove low temporal frequencies before fitting the GLM, and the first two volumes of each run (always extra null trials) were discarded to ensure data quality. Individual patterns for each run were normalized before cross-run comparison by calculating the *z*-score for each voxel, across conditions. *Z*-scored patterns were averaged within odd and within even runs of the same format (e.g., image runs 1 and 3 were averaged; video runs 2 and 4 were averaged) and discrimination index scores were

calculated based on correlations between even and odd sets of runs.

In order to produce optimal alignment of searchlight maps across subjects, we first reconstructed anatomical pial surface and gray-white matter boundaries for each subject using FreeSurfer v5.3.0 (*http://surfer.nmr.mgh. harvard.edu*; RRID:SCR_001847). These were aligned to a FreeSurfer standard template using a spherical transformation (Fischl, Sereno, Tootell, & Dale, 1999), and based on this alignment, the *mri_vol2vol* tool was used to calculate registrations from subject functional space to FreeSurfer standard. These standard-space subject maps were submitted to a second-level random-effects analysis in FSL. To correct for multiple comparisons, the group-level *t*-map was submitted to threshold-free cluster enhancement (TFCE; Smith & Nichols, 2009), an algorithm designed to offer the sensitivity benefits of cluster-based thresholding without the need for an arbitrarily chosen threshold. The TFCE statistic represents the cluster-like local support for each voxel using empirically and theoretically derived height and extent parameters. This TFCE map was then whole-brain corrected ($p < 0.05$) for the family-wise error rate using standard permutation tests implemented in FSL with the *randomise* function (10,000 permutations) and spatial 5-mm FWHM variance smoothing, which is recommended for $df < 20$ because it reduces noise from poorly estimated standard deviations in the permutation test procedure (Nichols & Holmes, 2002).

Searchlight analyses were also conducted within visual format (one for Image Format, one for Video Format). The same analyses as above were implemented, except for the following. For Image Format, patterns were compared between image runs only (e.g., *kicking* in the odd image runs with *kicking* in the even video runs); for Video Format, between video runs only (e.g., *kicking* in the odd video runs with *kicking* in the even image runs). To qualitatively compare the overlap of within- and cross-format decoding regions, we overlaid whole-brain searchlight maps for the different format comparisons to examine regions of conjunction. Here the maximum *p*-value (TFCE, whole-brain corrected) is the valid value for conjunction inference in each voxel (the Minimum Statistic compared to the Conjunction Null, MS/CN; Nichols, Brett, Andersson, Poline, & Wager, 2005).

2.5.4. Cross-Format ROI definition

We used the results of the cross-format searchlight analysis to define regions of

interest (ROIs) for three subsequent analyses, described below. ROIs were constructed by taking the intersection of the cross-format decoding map (whole-brain corrected) and spheres centered on the cluster peaks (Table 3.1) from this map (Fairhall & Caramazza, 2013), and transforming the defined region back into the native functional space for each subject. Since spheres with a given radius may yield different ROI sizes after intersection with the whole-brain map, the radius of these spheres was adjusted separately for each region so that approximately 100 voxels were contained within each ROI after transformation to subject space (mean 108 voxels, *sd* 15, range 81 to 156).

2.5.5. Invariance to controlled factors in the video stimuli

The first follow-up analysis tested whether the patterns elicited by the movies showed invariance to incidental properties of the actions, such as the action direction (leftward vs. rightward) and actor roles (actor A as Agent or actor B as Agent). To test whether this was the case, we implemented additional GLMs that included one regressor for each action category × action direction × actor role combination within each video run (32 regressors total per run, with 2 video stimuli contributing to each estimate). Multivoxel patterns within run were *z*-scored across the 32 conditions, and these patterns were averaged within odd and within even runs. For each cross-format ROI, pairwise Pearson correlations were calculated for patterns between all 32 conditions across odd and even runs, and correlation coefficients were averaged for all combinations of same vs. different action category, same vs. different action direction, and same vs. different actor roles, yielding eight mean correlations values per subject and ROI. These pattern similarity values were then entered into 2 × 2 × 2 repeated measures ANOVAs (one for each ROI), with action category, action direction, and actor roles as factors. Early visual cortex (EVC, defined in a separate functional localizer described below) was also included in this analysis for comparison with the cross-format ROIs. *P* values for *F* statistics were corrected for multiple comparisons across the nine ROIs using the Bonferroni-Holm method, separately for each set of *F* statistics yielded by the ANOVA. The Bonferroni-Holm method is uniformly more powerful than standard Bonferroni while still controlling for the family-wise error rate (Holm, 1979). Note that although the same Video Format data was used for cross-format ROI definition and for this follow-up analysis, this analysis is unlikely to be affected by circular analysis problems (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009), because the cross-format

ROI definition procedure used GLMs that collapsed across actor role and action direction for each action category. Thus, the Video patterns used in the cross-format ROI definition procedure did not contain information about actor role or action direction.

2.5.6. Representational Similarity Analysis

The second follow-up analysis tested whether the patterns that allow for action discrimination *within* individual reflect a common representational space *across* individuals, i.e., whether actions are represented in a similar way from person to person. To examine this issue, we used representational similarity analysis (Kriegeskorte & Kievit, 2013; RSA; Kriegeskorte, Mur, & Bandettini, 2008). Within each cross-format ROI, representational dissimilarity matrices (RDMs) were constructed using the pairwise Pearson correlation distances (1 - *r*) between multivoxel patterns for each action category to every other. Three separate RDMs were constructed for every subject and ROI: a video format RDM (even to odd video run correlations), an image format RDM (even to odd image run correlations), and a cross-format RDM (all video to all image run correlations). Cross-subject consistency in representational space was then assesses by calculating Spearman correlation between each subject's RDM and every other subject's RDM, separately for video, image, and cross-format RDMs. Because we were interested in similarities and differences between categories, rather than reliability within categories, only off-diagonal elements of the RDMs were included in the calculation. These inter-subject correlations represent the similarity in representational space from each subject to every other, abstracted away from the underlying units of representation (voxels). If the mean inter-subject correlation is significantly above zero, it indicates that the relationship among representational spaces is reliable.

Because the inter-subject RDM correlation values were not independent (i.e., RDMs from each subject were used more than once in the inter-subject RDM comparisons, e.g. subject 1 to subject 2, subject 1 to subject 3, etc.), permutation tests were used to determine chance levels. In these tests, for each comparison type (video, image, cross), the condition labels of the subject RDMs were shuffled before calculating the inter-subject correlations. The mean inter-subject correlation was calculated across all pairwise subject comparisons, 10,000 times for each comparison type and cross-format ROI. The *p* value was simply the proportion of times the true mean inter-subject correlation was lower than a permuted inter-subject correlation. These *p* values were

then corrected for multiple comparisons across the eight ROIs using the Bonferroni-Holm method, separately for each comparison type. The mean chance inter-subject correlation from permutation testing was approximately zero in all cases (mean 7.77×10^{-5}, range -1.50×10^{-3} to 1.50×10^{-3}, across all ROIs and comparison types).

Note that although the same data was used for cross-format ROI definition and for this follow-up analysis, the results of these analyses do not follow trivially from the finding of cross-format action category representations in these regions. In particular, since the action category discrimination index was quantified separately for each subject (using each subject's own representational space), reliable action category decoding across subjects does not logically entail that their representational spaces will be related to one another. To confirm this point, we ran a simulation using randomly constructed RDMs. We observed no correlation between the magnitude of the discrimination indices and the Spearman correlation of the off-diagonal values across RDMs (mean -6.48×10^{-4}, *sd* 0.03, across 1,000 simulations).

2.5.7. Functionally localized regions of interest

We also examined action decoding in several functional ROIs that previous work suggests might play a role in processing actions, or perceptual constituents of actions. These ROIs were defined based on fMRI responses during three functional localizer scans (described above).

Data from the first localizer scan were used to define ROIs related to the viewing of specific stimulus categories, using a group-constrained subject-specific (GSS) ROI definition method (Julian, Fedorenko, Webster, & Kanwisher, 2012). This approach yields similar individual subject functional ROIs to the traditional hand-drawn ROI pipeline, but using an objective and automatic method. Each ROI was initially defined in each subject as the top 100 voxels in each hemisphere that responded more to the contrast of interest and fell within the group-parcel mask for the given ROI. Parcel masks were derived from a large number of separate subjects undergoing similar localizers using this method (parcels available at *http://web.mit.edu/bcs/nklab/GSS.shtml*). Using this method, we identified the following ROIs, each using the contrast listed in parentheses: early visual cortex (EVC; scrambled objects > objects); lateral occipital (LO) and posterior fusiform (pFs; objects > scrambled objects); occipital face area (OFA), anterior fusiform face area (FFA), and

right posterior FFA (faces > objects); extrastriate body area (EBA) and right fusiform body area (FBA; bodies > objects); and occipital place area (OPA), parahippocampal place area (PPA), and retrosplenial complex (RSC; scenes > objects).

Data from the second localizer scan (dynamic stimuli) were used to define two motion-sensitive functional ROIs. GSS parcels were not available for these stimulus contrasts, so these ROIs were hand-drawn. Human middle temporal complex (hMT+) was defined as the set of contiguous voxels responding more to scrambled than static point-light displays in the vicinity of the posterior inferior temporal sulcus, separately in both hemispheres. Thresholds were determined separately for each subject to be consistent with ROIs found in previous studies (mean $t > 5.3$, range 3 to 8). The biological-motion selective posterior superior temporal sulcus (pSTS-bio) was defined as the set of contiguous voxels responding more to intact than scrambled point-light displays in the vicinity of the posterior superior temporal sulcus in the right hemisphere. Thresholds were determined separately for each subject to be consistent with ROIs found in previous studies (mean $t > 2.9$, range 2.0 to 4.7, identified in 11 of 15 participants).

Data from the third localizer scan (linguistic stimuli) were used to define the verb-selective left pMTG (pMTG-verb) as the set of contiguous voxels responding more to verbs than nouns in the vicinity of the left posterior middle temporal gyrus. Thresholds were determined separately for each subject to be consistent with ROIs found in previous studies (mean $t > 3.7$, range 2.4 to 4.5, identified in 11 of 15 participants).

Because these functional ROIs often partially overlapped in individual subjects, we excluded voxels falling into more than one ROI (Schwarzlose, Swisher, Dang, & Kanwisher, 2008; Weiner & Grill-Spector, 2013). This allowed us to isolate the specific contribution of voxels with certain functional profiles (e.g., body-selective or motion-selective), without contamination from nearby regions with different functional profiles. After these exclusions, the mean size of the ROIs was as follows: EVC: 186 voxels (*sd* 15, range 150 to 200); hMT+: 146 voxels (*sd* 30, range 98 to 220); pSTS-bio: 51 voxels (*sd* 22, range 16 to 93); LO: 155 voxels (*sd* 15, range 134 to 172); pFs: 142 voxels (*sd* 21, range 86 to 163); anterior FFA: 150 voxels (*sd* 17, range 122 to 178); right posterior FFA: 200 voxels (no overlap); OFA: 165 voxels (*sd* 23, range 114 to 193); EBA: 116 voxels (*sd* 24, range 87 to 160); right FBA: 65 voxels (*sd* 13, range 43 to 92); OPA: 195 voxels (*sd* 3,

range 190 to 200); PPA: 181 voxels (*sd* 14, range 147 to 197); RSC: 200 voxels (no overlap); pMTG-verb: 94 voxels (*sd* 60, range 35 to 211). Analyses using ROIs in which overlapping voxels were not excluded yielded qualitatively similar results.

Action category discrimination indices for the video, image, and cross-format comparisons were calculated separately within each ROI for each subject, and were submitted to two-tailed one-sample *t* tests against chance (zero). *P* values were corrected for multiple comparisons across functional ROIs separately within comparison type using the Bonferroni-Holm method (14 tests for each comparison type).

2.6. Eyetracking control task

To ensure that action category decoding could not be attributed to differences in spatial attentional allocation, we ran a control study in which a separate group of participants underwent the identical procedure as in the main fMRI experiment, but outside the scanner, and while their gaze location was recorded by a remote binocular eyetracker situated within the visual display monitor (Tobii T120 eyetracker sampling at 60 Hz).

Two-dimensional gaze maps were created for each combination of subject, format (Image or Video), run (4 per format), and action category (8) by binning gaze locations on the screen into 70 horizontal × 56 vertical bins. In other words, gaze maps akin to a two-dimensional histogram were formed by dividing the screen extent into 70 × 56 bins, and each eyetracking sample was placed into its corresponding location in this set of bins (ignoring the time dimension). As with the fMRI voxel patterns, these gaze maps were *z*-scored across action category (for each subject, format, and run), and even and odd run maps were averaged together. We then attempted to decode action category both within- and across-format using the two-dimensional gaze maps. Pearson correlations were calculated between even- and odd-run gaze maps corresponding to each action category (for each subject and analysis type separately). The discrimination index was the average within-category correlation minus the average between-category correlation. We tested the significance of this discrimination index across subjects, separately for Image Format, Video Format, and Cross Format.

3. Results

3.1. Behavioral performance

One participant reported that she misunderstood the instructions for her first video run, so data for this run (behavioral and imaging) were excluded. For the remaining data, behavioral performance on the one-back repetition detection task was good, indicating that participants were playing close attention to the stimuli. For image runs, the mean accuracy on repetition trials was 0.91 (*sd* 0.08), the mean false alarm rate was 0.002 (*sd* 0.002), and average RT on correct trials was 694 ms (*sd* 82 ms). For video runs, mean accuracy was 0.89 (*sd* 0.10), the mean false alarm rate was 0.014 (*sd* 0.015), and average RT on correct trials was 1,117 ms (*sd* 157 ms).

3.2. Cross-format action category decoding across the brain

Our primary goal was to identify representations of action categories that were invariant to incidental visual elements, such as actors, objects, scene context, and the presence or absence of dynamic motion information. To this end, we scanned participants while they viewed videos and still images of eight categories of interactions. We then used a searchlight analysis to identify brain regions where action category representations could be decoded across the video and image formats. This analysis revealed seven contiguous clusters in the cross-format searchlight volume, which were located in left and right inferior parietal lobule (IPL), left and right lateral occipitotemporal cortex (LOTC), left and right ventral occipitotemporal cortex (VOTC), and left middle frontal gyrus (mFG; see Figure 3.2*A*, and see Table 3.1 for list of these clusters). These regions largely overlap with the previously identified Action Observation Network (AON; Caspers et al., 2010; Kilner, 2011; Rizzolatti & Sinigaglia, 2010; Urgesi et al., 2014). These results suggest that AON regions encode categories of actions in a consistent way across highly varied perceptual input. For subsequent analyses, ROIs corresponding to these clusters (approximately 100 voxels each) were defined individually in each subject. (For discussion of the relationship of the cross-format OTC regions to functionally defined OTC regions based on previous literature, see below, Cross-format decoding in functionally selective regions.)

The largest cluster, left IPL, had several local maxima (Table 3.1). The cluster peak was in left ventral IPL in the supramarginal gyrus (xyz_{mni} = -58, -37, 28), and this was

used as the left IPL ROI for further analyses. An additional local maximum was located in left premotor cortex (xyz_{mni} = -55, -4, 40; Figure 3.2*A*). Though this area was contiguous with the left IPL cluster in the volume, it is anatomically separated by several sulci and gyri from the other local maxima, and prior literature suggests a possible functionally distinct role for left premotor cortex in recognition of actions (Caramazza et al., 2014; Kilner, 2011; Rizzolatti & Sinigaglia, 2010; M. F. Wurm & Lingnau, 2015). Thus, we defined an additional ROI around this local maximum for further interrogation. With this additional ROI, we had eight ROIs for subsequent analyses: left and right IPL, left and right LOTC, left and right VOTC, left premotor, and left mFG.

Table 3.1

Cluster						Peak					
Region	Extent	Cross-format discrim. index	*x, y, z* (Center of Gravity)			$p_{(FWE-cor)}$	Pseudo-*t*	$p_{(unc)}$	x, y, z		
Left IPL (ventral) †	605	0.045	-49.2	-26.7	38.7	0.002	5.07	1E-04	-58	-37	28
Left Premotor (ventral) ††						0.003	3.85	9E-04	-55	-4	40
Left IPL (dorsal)						0.02	5.73	3E-04	-40	-43	49
Left Post-Central						0.003	5.03	3E-04	-58	-22	37
Left Premotor (dorsal)						0.031	3.56	5E-04	-31	-4	43
Right LOTC †	96	0.049	48.9	-61.5	6.1	0.004	6.11	2E-04	44	-61	4
Right IPL †	64	0.045	56.3	-27.9	37.3	0.016	4.91	3E-04	53	-28	37
Left LOTC †	28	0.050	-42.9	-80.6	-0.1	0.026	4.54	1E-03	-43	-82	-2
Left VOTC †	25	0.043	-32.8	-44.8	-12.6	0.019	5.09	5E-04	-28	-40	-11
Right VOTC †	17	0.053	44.4	-52.0	-12.9	0.029	4.88	9E-04	44	-52	-11
Left mFG †	11	0.033	-45.5	20.8	30.2	0.036	3.85	2E-04	-46	23	28

MNI locations, extent, mean cross-format discrimination index, significance, and peak statistics for the clusters identified in the cross-format action category searchlight, ordered by cluster extent (number of voxels). Indented are MNI locations and statistics for peaks of additional local maxima within these clusters that were separated by at least 15 mm in the volume. The ROIs used in subsequent analyses were composed of approximately 100 voxels centered on the cross-format cluster peaks (marked with † symbol), with the addition of the local maximum for left premotor (ventral only, marked with †† symbol). Abbreviations: cor, corrected; IPL, inferior parietal lobe; LOTC, lateral occipitotemporal cortex; mFG, middle frontal gyrus; unc, uncorrected; VOTC, ventral occipitotemporal cortex.

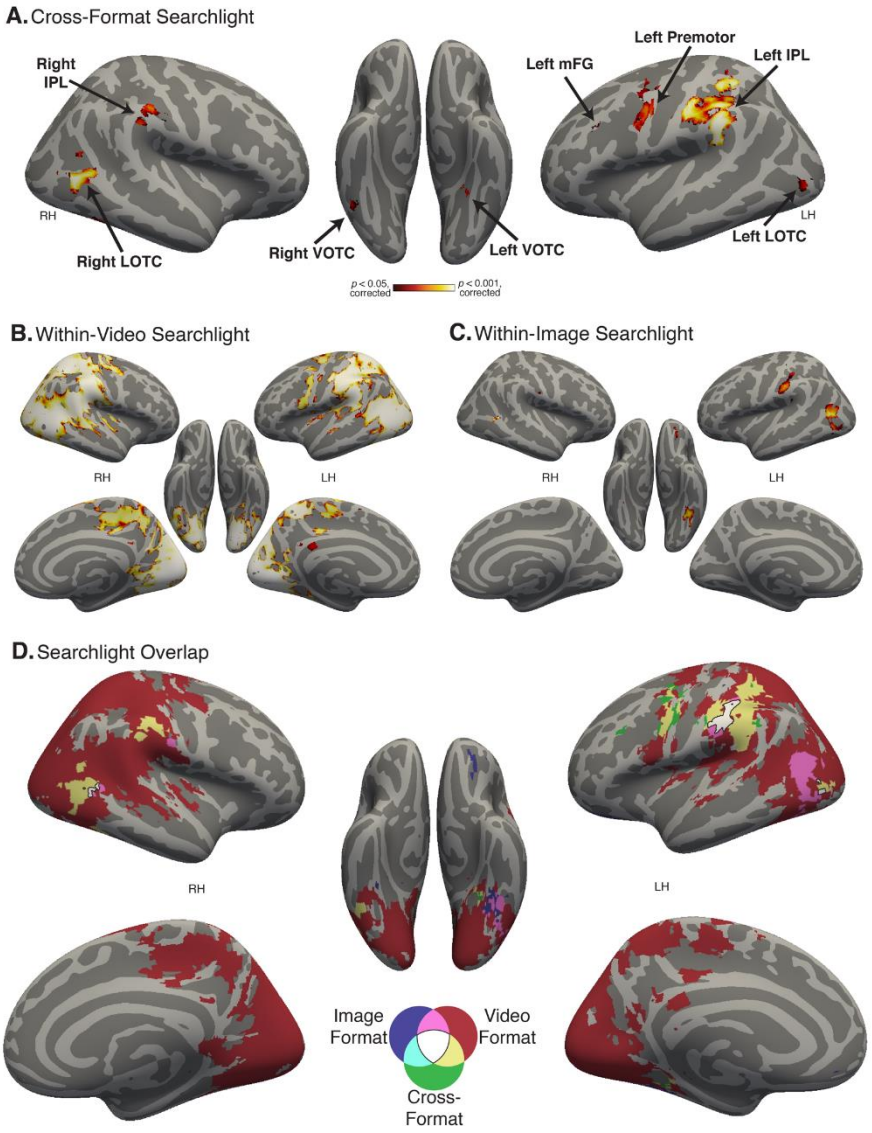


Figure 3.2.

A, Whole brain searchlight for Cross-Format action category decoding. Black arrows and text indicate the anatomical locations of the cross-format clusters identified in this analysis, as well as the location of the ROI for left premotor cortex. Corrected for multiple comparisons at $p < 0.05$, using threshold-free cluster enhancement (TFCE) and permutation testing. For subsequent analyses, ROIs corresponding to these clusters were defined individually in each subject. ***B,*** Whole brain searchlight for Image Format action category decoding (corrected as in ***A***). ***C,*** Whole brain searchlight for Video Format action category decoding (corrected as in ***A***). ***D,*** Conjunction overlap map of the searchlight analyses, with colors indicating which of

the three searchlight maps overlap in each area (black outline indicates overlap of all three). Video Format decoding was widespread across the brain, while Image Format decoding was mostly restricted to similar regions as were found in the Cross-Format searchlight.

Prior work has shown coding of specific limb and effector information in some of the regions reported here (Bracci & Peelen, 2013; J. P. Gallivan, McLean, Smith, & Culham, 2011; Jason P. Gallivan, Adam McLean, Valyear, & Culham, 2013; IPL and LOTC; Mahon et al., 2007; Orlov, Makin, & Zohary, 2010; Orlov, Porat, Makin, & Zohary, 2014; R. Peeters et al., 2009; R. R. Peeters, Rizzolatti, & Orban, 2013). To ensure that the present cross-decoding results were not driven solely by an effector-based distinction between action categories, we examined cross-format decoding separately for two sets of our action categories: those that involved hand/arm effectors (*massaging, pulling, shoving, slapping, tapping*), vs. those that involved other, more varied, effectors (*biting, brushing, kicking*). If an effector-based distinction between hand/arm and non-hand/arm actions were driving our results, we should observe cross-format decoding within the varied effector set and not the hand/arm effector set. However, despite the reduced data available in each subset, we still observed cross-decoding in half or more of the cross-format ROIs in both subsets: Six of eight ROIs showed significant decoding within the varied effector set (left mFG, left VOTC, left and right LOTC, left and right IPL; t values > 2.75, $p_{\text{corrected}}$ values < 0.046), and four of eight showed significant decoding within the hand/arm effector set (right VOTC, right LOTC, left and right IPL; t values > 3.02, $p_{\text{corrected}}$ values < 0.046). This suggests that, in these regions at least, cross-format decoding is unlikely to be driven solely by a coarse distinction between actions performed with the hand/arm vs. other effectors.

3.3. Within-format action category decoding across the brain

To determine whether action category information tied to the particular visual format was present in other brain regions, we conducted whole-brain searchlights for action category decoding, separately for the video format and image format. Within the video format, we found widespread action category decoding across the brain in both hemispheres (Figure 3.2*B*). These results are not surprising, given the consistency in visual motion energy across the video clips within action category (see above, Methods: Video stimuli). In contrast, action category decoding within the image format was restricted largely to the regions identified in cross-format decoding (Figure 3.2*A* vs.

Figure 3.2*C*), with an additional left orbitofrontal cluster. This was confirmed in a conjunction overlap map of the three searchlight maps (Figure 3.2*D*; Nichols et al., 2005): The within-format searchlights overlapped one another in or adjacent to areas observed in the cross-format searchlight. Interestingly, the degree of overlap of the maps in the different regions suggests a possible difference in the degree of format dependence of action coding between left IPL and the other regions. In the former, there is a large area of cross-decoding, and the within-format territory (both image and video) overlaps with this. In the other regions, particularly left LOTC, there is only a small area of cross-decoding, but large (and overlapping) areas of within-format decoding. This might suggest that action representations are less format-dependent in left IPL than in other regions.

3.4. Can the cross-format results be driven by similarities in spatial location of attention?

The cross-format results might have been trivially obtained if participants attended to similar spatial locations for each action category, even across the two visual formats (image and video). For example, it is reasonable to hypothesize that for *kicking*, participants might have attended to the lower portion of the visual display, whereas for *slapping*, they attended to the higher portion. Such consistency in location of spatial attention has been shown to drive multivoxel responses in visual regions, including hMT+ (Bressler & Silver, 2010; Connor et al., 2002). In order to rule out this possibility, we conducted a control study, in which a separate group of 16 participants performed the same task as the fMRI participants while their gaze was tracked with a remote eyetracker. These gaze data were analyzed similarly to how the fMRI data were analyzed, i.e. multivariate patterns (here, two-dimensional maps of gaze location) were constructed for each subject, format, and run, and discrimination indices were calculated from the correlation of the 2-D gaze maps across action categories for each subject and format. If participants looked to consistent spatial locations for each action category, these gaze map discrimination indices would be reliably above zero.

Action category could indeed be decoded based on gaze location for both the Image Format, $t_{(15)} = 2.60$, $p = 0.02$, and the Video format, $t_{(15)} = 7.91$, $p < 0.001$. However, across the visual formats, discrimination indices based on gaze locations were reliably

below zero, $t_{(15)} = -4.26$, $p < 0.001$. These results indicate that gaze locations for action categories were consistent within format, but were systematically different across formats (e.g., looking at the top half of the screen for *kicking* in the Image Format, but the lower half for *kicking* in the Video Format). Thus, absolute location of spatial attention is unlikely to explain the cross-format decoding results from fMRI data in the main experiment.

3.5. Invariance to systematically manipulated properties of the video stimuli

Abstract action category representations should show generalization not only across formats, but also across variations in incidental properties within format, such as actors or viewpoint/action direction. Some evidence that this may be the case comes from the fact that we were able to decode action category using only patterns elicited by the images, even though the image stimuli were chosen to maximize within-category visual dissimilarity. However, to formally test generalization across incidental properties, we leveraged the fact that actor roles and action direction were systematically manipulated in the video clips. We extracted activity patterns within each ROI for each specific condition (i.e., each action category × action direction × actor role combination, 32 patterns in total). The correlation values between these conditions were then calculated and entered into repeated measures ANOVAs (one for each ROI), with action category (same vs. different), action direction (same vs. different), and actor roles (same vs. different) as factors. We also included early visual cortex (EVC) in the analysis (defined in a separate functional localizer as responses to scrambled objects > intact objects) as an indicator of whether it was possible to detect differences in action decoding across incidental low-level visual properties in our data.

Finding action category decoding was expected in this analysis, as the ROIs were selected based on the presence of consistent action category patterns across format, which entails that the patterns within the Video Format should also be consistent. Somewhat surprisingly, action category decoding was robust in some but not all regions (Figure 3.3*A*). This might be attributable to more variability in the estimates of activity patterns (there were two trials per beta estimate in the GLM used here, as opposed to eight trials per beta estimate in the previous analysis). Nevertheless, the estimates were

consistent enough that seven of the eight cross-format ROIs, plus EVC, showed either a main effect of action category or a trend in this direction. These effects were marginal in left mFG, left premotor, and right VOTC ($F_{(1,14)}$ values of 4.66, 5.91, and 5.63, $p_{\text{corrected}}$ values = 0.12, $p_{\text{uncorrected}}$ values < 0.05), and significant in all other regions ($F_{(1,14)}$ values > 21.9, $p_{\text{corrected}}$ values < 0.002, $p_{\text{uncorrected}}$ values < 0.001) except left VOTC ($F_{(1,14)}$ = 0.75, $p_{\text{corrected}}$ = 0.40, $p_{\text{uncorrected}}$ = 0.40). For action direction, a subset of regions showed main effects (EVC and left LOTC significant, right LOTC marginal), with greater pattern similarity for the same action direction than different action direction (EVC: $F_{(1,14)}$ = 17.3, $p_{\text{corrected}}$ = 0.009, $p_{\text{uncorrected}}$ = 0.001; Left LOTC: $F_{(1,14)}$ = 10.3, $p_{\text{corrected}}$ = 0.05, $p_{\text{uncorrected}}$ = 0.006; Right LOTC: $F_{(1,14)}$ = 4.40, $p_{\text{corrected}}$ = 0.38, $p_{\text{uncorrected}}$ = 0.055; all other $F_{(1,14)}$ values < 3.03, $p_{\text{corrected}}$ values > 0.62, $p_{\text{uncorrected}}$ values > 0.10; Figure 3.3*A*). This suggests that these regions are sensitive to the direction of motion in the videos, which is not surprising, given the presence of motion-selective regions (hMT+) in LOTC, and EVC's role in coding low-level visual features. No ROI showed a main effect of actor roles (all $F_{(1,14)}$ values < 2.53, $p_{\text{corrected}}$ values > 0.99, $p_{\text{uncorrected}}$ values > 0.13; Figure 3.3*A*), indicating that no region distinguished videos with Actor A as the agent from videos with Actor B as the agent.

Crucially, in terms of action category invariance, no cross-format ROI showed an interaction of action category with actor role and/or action direction; if anything, in left premotor cortex, action decoding was marginally better for *different* action direction vs. same ($F_{(1,14)}$ = 4.07, $p_{\text{corrected}}$ = 0.51, $p_{\text{uncorrected}}$ = 0.06; all other cross-format ROI $F_{(1,14)}$ values < 2.91, $p_{\text{corrected}}$ values > 0.99, $p_{\text{uncorrected}}$ values > 0.11; see Figure 3.3*B*). While the lack of significant interactions is a null result and should be interpreted with caution, it is worthwhile to note that this modulation *was* detectable in our data: in EVC, action categories could be better decoded when action directions were the same than when they were different (action category × action direction interaction: $F_{(1,14)}$ = 12.4, $p_{\text{corrected}}$ = 0.03, $p_{\text{uncorrected}}$ = 0.003; Figure 3.3*B*). Thus, in regions showing cross-decoding of action category across videos and images, the ability to distinguish action categories was no greater when comparing across patterns elicited by videos in which actor role or action direction were the same than when comparing across videos in which actor role or action direction were different. Although we cannot definitively rule out the possibility that action representations in these cross-format regions are modulated by visual properties

of the video stimuli, this finding is at least consistent with abstract action category codes.

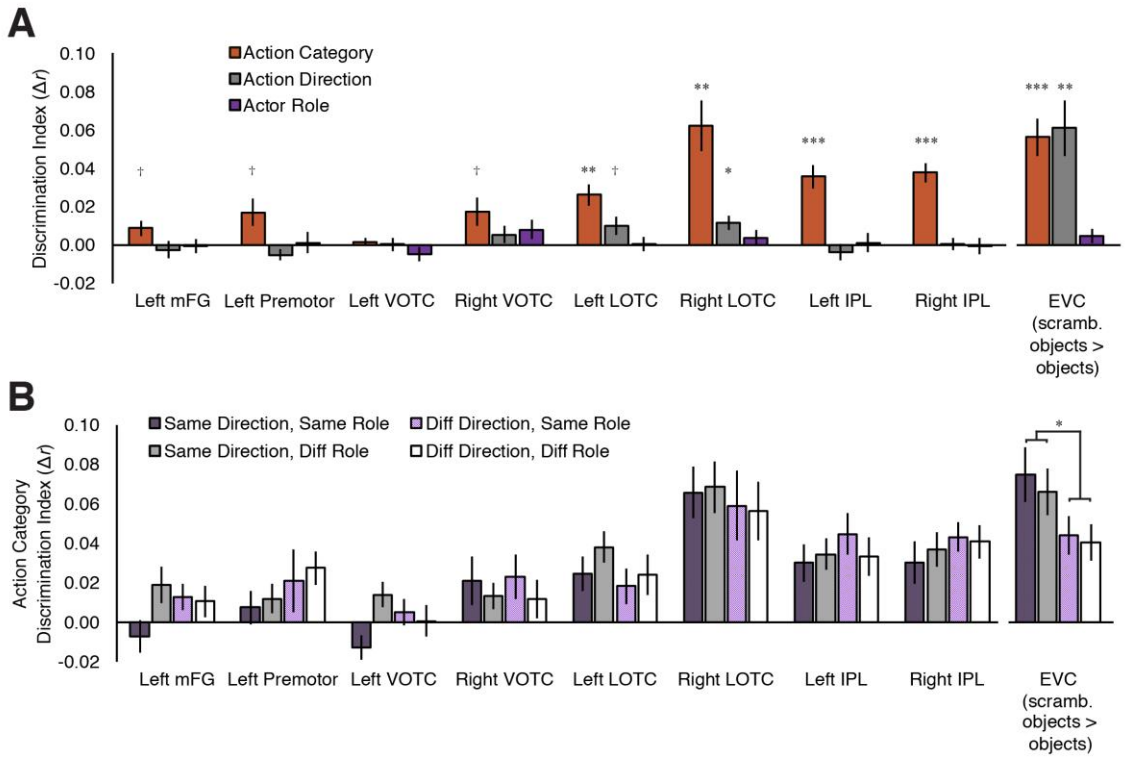


Figure 3.3

Analyses for action category specificity and generalization for the Video Format stimuli in cross-format ROIs. Early Visual Cortex (EVC, defined by a functional localizer as scrambled objects > intact objects) was also included for comparison with cross-format ROIs. ***A,*** Decoding for action category, action direction, and actor roles. Discrimination index values shown here are average same minus average different correlation values for action category, action direction, and actor roles, respectively, collapsed over the other factors. Action direction could be decoded in left and right LOTC and EVC, while actor role code not be decoded in any regions. Action category could be decoded in most regions, though we note that this is necessitated by our ROI selection procedure, which was based on cross-format action category decoding using the same data. ***B,*** Action category discrimination indices for the cross-format ROIs, for each combination of action direction (same or different) and actor roles (same or different), i.e., the orange Action Category bars in ***A*** split by the other factors. Only significant differences between action category decoding are indicated. Action category representations were largely invariant to the systematically manipulated properties of the video stimuli in cross-format ROIs, while in EVC, action category decoding was significantly better when action direction was the same vs. different. † $p < 0.055$, uncorrected; * $p < 0.05$; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple comparisons across the nine ROIs. Error bars represent SEMs.

3.6. Representational Similarity Analysis

Although recognizing actions as distinct from one another is a crucial first step towards action understanding, reasoning and communicating about actions requires more graded appreciation of similarities and differences between action categories. For

example, two people may readily distinguish *slapping* from *shoving*, corresponding to successful action recognition for each individual. But if two people's representational spaces further indicate that *slapping* is very similar to *shoving*, mutual understanding and communication about these actions will be facilitated. To determine the extent to which representational spaces for action categories were consistent across individuals, we calculated the Spearman correlation between off-diagonal values of representational dissimilarity matrices (RDMs) for each subject to every other subject, separately for each cross-format ROI and comparison type (image format, video format, and cross-format). The mean inter-subject correlation is the average consistency in representational spaces across individuals, where chance is zero.

For the image format comparisons, no ROI showed significant consistency in representational space across subjects ($p_{\text{corrected}}$ range 0.07 to 0.49), although four of the eight showed consistency uncorrected for multiple comparisons (left and right VOTC, right LOTC, and right IPL, $p_{\text{uncorrected}} < 0.05$; other $p_{\text{uncorrected}}$ values > 0.10). In contrast, for the video format, six of eight ROIs showed consistency across subjects (left premotor, right VOTC, left and right LOTC, and left and right IPL, $p_{\text{corrected}}$ values < 0.05, $p_{\text{uncorrected}}$ values < 0.005; other $p_{\text{corrected}}$ values > 0.18, $p_{\text{uncorrected}}$ values > 0.09). Similar findings to the video format were obtained for cross-format consistency: the same six ROIs showed consistency across subjects ($p_{\text{corrected}}$ values < 0.03, $p_{\text{uncorrected}}$ values < 0.009; other $p_{\text{corrected}}$ values > 0.82, $p_{\text{uncorrected}}$ values > 0.41). (Inter-subject correlation values are depicted in Figure 3.4*A*; see Figure 3.4*B* for a visualization of the clustering of action categories across regions.)

It is at first glance puzzling that the cross-format consistency was reliable in most regions, despite the lower image format consistency. One account of these contrasting results appeals to the difference in reliability of the "action category signal" between the image and video formats, which should be greater for the video format (as indicated by the higher norming label agreement in this format). For cross-format consistency, the robust video format action category signal may "pull out" the weaker image format signal, even when the comparison is made across subjects. The plausibility of this account was confirmed by simulations. We generated sets of action category signals with different levels of signal-to-noise (SNR), and compared the resulting inter-subject consistencies. Specifically, we generated one "true" activity pattern for each of eight

action categories made up of 100 voxel responses randomly drawn from a Gaussian distribution $N(0,1)$. Varying degrees of noise were added to these "true" underlying action category patterns, separately for 8 runs (4 for each visual format) for each of 15 simulated subjects. This noise was systematically varied for each format by choosing standard deviations from the set (0.01, 0.10, 0.50, 1, 3), separately for the video and image formats (i.e., the video format standard deviation might be 0.10 while the image format standard deviation might be 3). The same RSA as described above was then conducted using these simulated activity patterns. These simulations revealed that comparisons of RDMs built from two sets of low-SNR action category patterns (equivalent to the image-format comparison, in this account) show a much less consistent relationship than comparisons of RDMs built from one high- and one low-SNR set of action category patterns (the cross-format comparison, in this account). Together, these analyses suggest that most regions we have identified contain a representational space that generalizes from person to person, even when this space is built from two different visual formats.

[Manuscript continues with figure on next page]

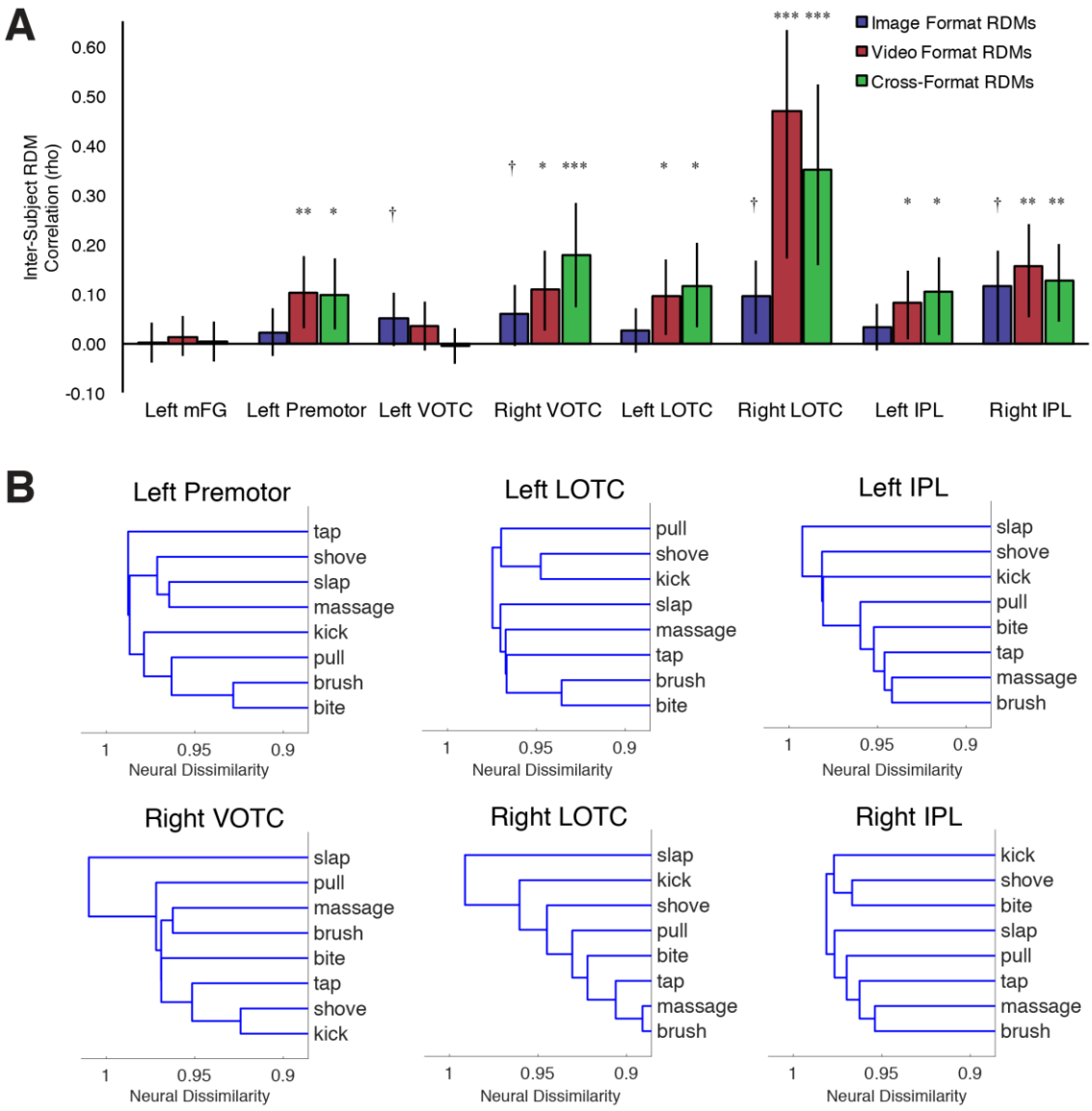


Figure 3.4

Cross-subject Representational Similarity Analysis. RDMs (representational dissimilarity matrices) for each subject were constructed from the multivoxel patterns for each action category and were compared across subjects. ***A***, Mean inter-subject RDM correlation across all pairwise comparisons of the 15 subjects, separately for the image RDMs, video RDMs, and cross-format RDMs. Representational spaces for action categories were consistent across both subjects and formats in bilateral LOTC, bilateral IPL, right VOTC, and left premotor. † $p < 0.05$, uncorrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple comparisons across the eight ROIs (separately for each comparison type). Permutation tests were used to determine significance, based on a null distribution of the correlation statistic generated from 10,000 random permutations in which action category labels were shuffled before calculation of the RDM correlations. *P* values are the proportion of permutation correlation statistics that were greater than the true statistic. Error bars here indicate the spread of the null distribution (95% of the null distribution width), centered at the mean inter-subject RDM correlation value. ***B***, Dendrograms depicting the hierarchical clustering of action categories in each cross-format ROI that showed significant cross-format and cross-

subject consistency, based on mean cross-format RDMs across subjects. Neural dissimilarity is displayed in Pearson correlation distance (1 - *r*). Distances between clusters were computed using the Matlab *linkage* function with the average distance algorithm.

3.7. Action category decoding in functionally selective regions

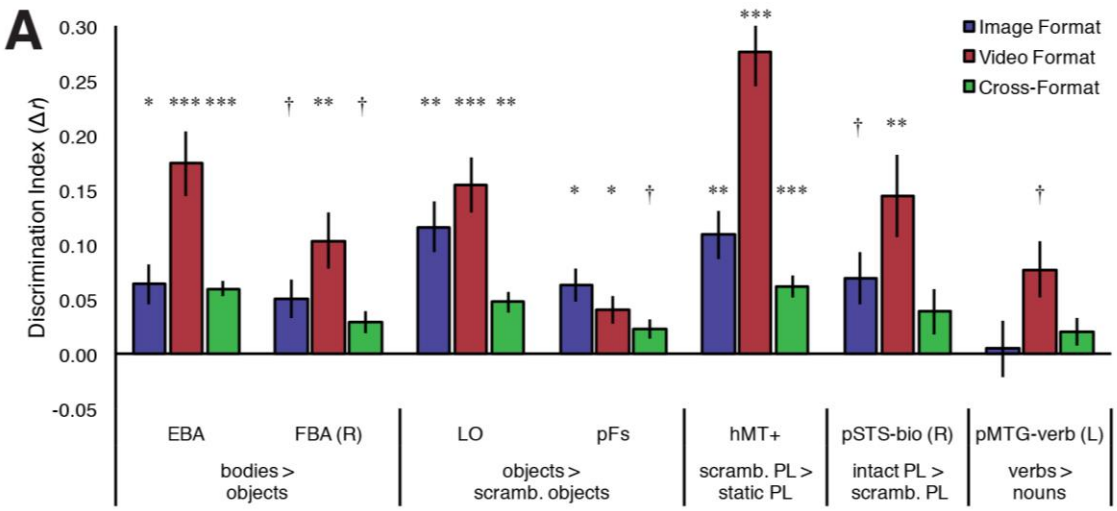
Although we focus above on regions identified in a hypothesis-free searchlight analysis, there are several well-studied functional regions (fROIs) in or near to occipitotemporal cortex that one might postulate a priori should have a role in action perception. These include motion-selective hMT+, body-selective EBA and FBA, object-selective LO and pFs, and biological motion-selective pSTS (Grill-Spector & Weiner, 2014; E. D. Grossman & Blake, 2002; Kanwisher, 2010; Kourtzi & Kanwisher, 2001; Peelen, Wiggett, & Downing, 2006; Tootell et al., 1995). Additionally, a region in LOTC just anterior to hMT+, the left posterior middle temporal gyrus (pMTG), has been found to respond to linguistic descriptions of actions (Bedny et al., 2008; Bedny, Caramazza, Pascual-Leone, & Saxe, 2012; Bedny et al., 2014; Peelen, Romagno, & Caramazza, 2012) and to respond in action tasks involving both words and static images (C. Watson, Cardillo, Ianni, & Chatterjee, 2013). To test the possibility that some of these regions might support abstract action representations, we performed the cross-format decoding analysis in these fROIs (see also Jason P. Gallivan, Chapman, Mclean, Flanagan, & Culham, 2013; Jason P Gallivan & Culham, 2015). As a control, we also examined other fROIs that we did not expect to be involved in abstract action category representations (Epstein & Kanwisher, 1998; Kanwisher, 2010; face- and scene-selective regions, and early visual cortex; Kanwisher, McDermott, & Chun, 1997).

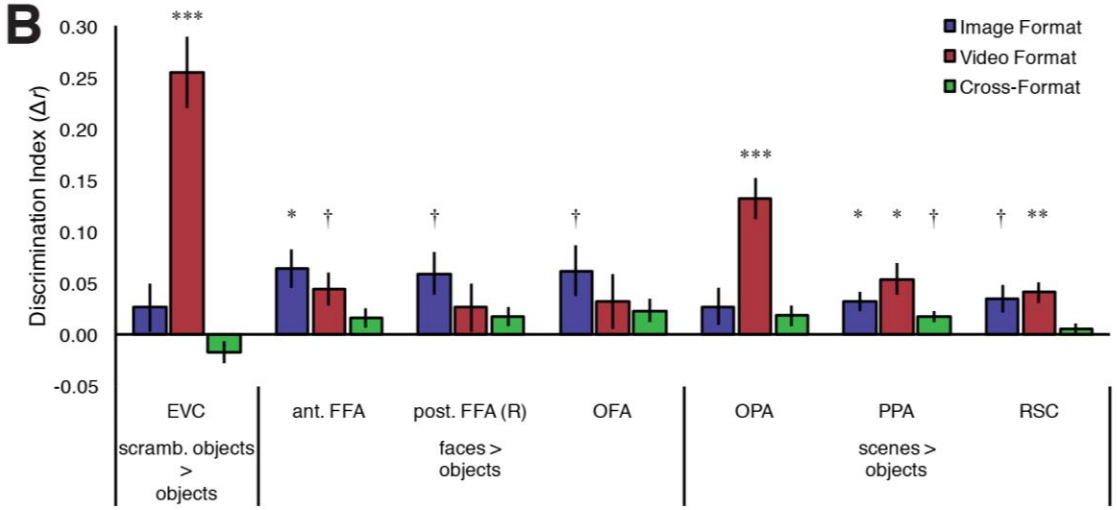
The only fROIs tested in which significant cross-format decoding was found were EBA, LO, and hMT+ ($t_{(14)}$ values > 5.20, $p_{\text{corrected}}$ values < 0.002, $p_{\text{uncorrected}}$ values < 0.001; Figure 3.5*A* and 3.5*B*). We did not find evidence for reliable cross-format decoding in other regions, though FBA, pFs, and PPA showed cross-format decoding at an uncorrected level (t values of 2.98, 2.78, and 3.02, $p_{\text{corrected}}$ values < 0.13, $p_{\text{uncorrected}}$ values < 0.02; all other t values < 2.13, $p_{\text{corrected}}$ values > 0.41, $p_{\text{uncorrected}}$ values > 0.052). Notably, we did not find clear evidence for cross-format action category decoding in two regions known to code for action-relevant stimuli: the biological motion-selective right pSTS ($t_{(10)} = 1.92$, $p_{\text{corrected}} = 0.59$, $p_{\text{uncorrected}} = 0.08$) and the verb-selective left pMTG ($t_{(10)} = 1.66$, $p_{\text{corrected}} = 0.59$, $p_{\text{uncorrected}} = 0.13$). Taken together, these results suggest that the

EBA, LO, and hMT+ are not only involved in representing bodies, objects, and motion, but also contribute to analysis of visual action scenes at an abstract level. (For a qualitative sense of the spatial relationship of fROIs and cross-format searchlight decoding, see Figure 3.5*C*.)

Besides the cross-format results in EBA, hMT+, and LO, several fROIs were sensitive to the action category information depicted within only the video format or image format (Figure 3.5*A* and 3.5*B*). However, the fact that these fROIs did not demonstrate cross-format decoding suggests that their role in representing actions at an abstract level is limited. Additionally, the absence of within-image format decoding in early visual cortex suggests that we adequately varied low-level image properties within action category.

[Manuscript continues with figure on next page]


A
Discrimination Index (Δr)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Image Format
Video Format
Cross-Format
EBA
FBA (R)
LO
pFs
hMT+
pSTS-bio (R)
pMTG-verb (L)
bodies > objects
objects > scramb. objects
scramb. PL > static PL
intact PL > scramb. PL
verbs > nouns


B
Discrimination Index (Δr)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Image Format
Video Format
Cross-Format
EVC
ant. FFA
post. FFA (R)
OFA
OPA
PPA
RSC
scramb. objects > objects
faces > objects
scenes > objects

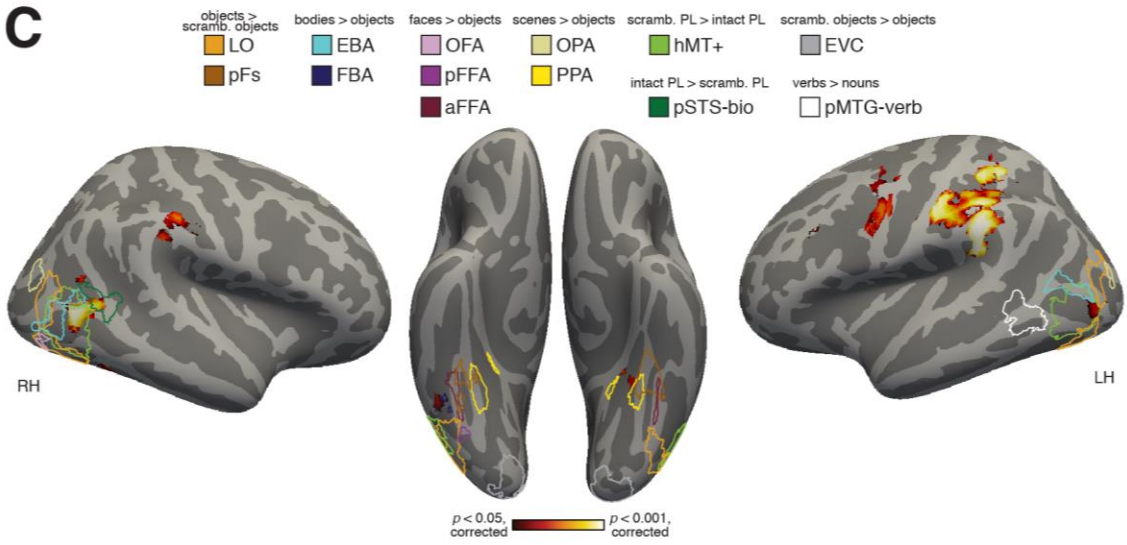

C
objects > scramb. objects
LO
pFs
bodies > objects
EBA
FBA
faces > objects
OFA
pFFA
aFFA
scenes > objects
OPA
PPA
scramb. PL > intact PL
hMT+
intact PL > scramb. PL
pSTS-bio
scramb. objects > objects
EVC
verbs > nouns
pMTG-verb
RH
LH
p < 0.05, corrected
p < 0.001, corrected

Figure 3.5
Action category discrimination indices for functionally defined regions (fROIs), for each comparison type (within-image format, within-video format, and cross-format). The only fROIs tested in which significant cross-format decoding was found were EBA, LO, and hMT+. ***A,*** Functional regions predicted to be sensitive to action category across format. ***B,*** Functional regions predicted to show minimal sensitivity to action category across format. Listed below each fROI is the localizer contrast used to define the region (e.g., bodies > objects). † $p < 0.05$, uncorrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple comparisons across the 14 fROIs (separately for each comparison type). Error bars represent SEMs. ***C,*** Visualization of the locations of fROIs relative to brain regions in which cross-format action category decoding was found. Cross-decoding searchlight map is identical to that in Figure 3.2*A* (i.e., corrected for multiple comparisons at $p < 0.05$). Outlines of fROIs were created by transforming individual subjects' fROIs to standard space and computing a group *t* statistic. Group fROIs were thresholded at $p < 0.001$ (uncorrected) except for the following fROIs, for which lower thresholds were needed for visualization: Left OPA, Right OFA, and pMTG-verb ($p < 0.01$); Left EBA ($p < 0.05$); and Left ant. FFA, Right FBA, and Right post. FFA ($p < 0.33$). RSC is not shown because no significant cross-decoding appeared on medial surfaces. Abbreviations: ant., anterior; EBA, extrastriate body area; EVC, early visual cortex; FBA, fusiform body area; FFA, fusiform face area (aFFA and pFFA: anterior and posterior FFA, respectively); hMT+, human middle temporal complex; LO, lateral occipital; OPA, occipital place area; pFs, posterior fusiform; PL, point-light; pMTG-verb, verb-selective posterior middle temporal gyrus (left only); post., posterior; PPA, parahippocampal place area; pSTS-bio, biological motion-selective posterior superior temporal sulcus (right only); (R), right; RSC, retrosplenial complex; scramb., scrambled.

4. Discussion

The goal of this study was to identify brain regions that mediate visual recognition of actions. We posited that these regions should display three key properties. First, they should support representations that discriminate between action categories, but are at least partially invariant to incidental features such as actor role, scene background, or viewpoint. Second, these action representations should be elicitable by both dynamic and static perceptual input. Third, these regions should not only discriminate hand-object interactions, but whole-body interactions with different effectors. By utilizing cross-format decoding methods, we identified several regions with these properties: bilateral OTC (lateral and ventral), bilateral IPL, left premotor cortex, and left mFG. The subset of these regions previously identified as the AON (LOTC, IPL, and left premotor; Caspers et al., 2010; Kilner, 2011; Rizzolatti & Sinigaglia, 2010; Urgesi et al., 2014) also exhibited consistency in representational space across subjects, a property that can facilitate a common understanding of actions among individuals.

Our findings add to the growing evidence that LOTC is involved in the coding of action categories (Oosterhof et al., 2010, 2012a, 2012b; Gallivan et al., 2013b; Watson et al., 2013; Gallivan and Culham, 2015; Tarhan et al., 2015; Tucciarelli et al., 2015; Wurm and Lingnau, 2015; Wurm et al., 2015; for review, see Lingnau and Downing, 2015). In

particular, our analyses of functional ROIs indicated that areas in LOTC selective for bodies, objects, and motion are also involved in visual action recognition from varied perceptual input: cross-format action category decoding was observed in EBA, LO, and hMT+ (see above, Action category decoding in functionally selective regions; Downing et al., 2001; S Ferri, Kolster, Jastorff, & Orban, 2013; Kourtzi & Kanwisher, 2001; for review, see Lingnau & Downing, 2015; Peelen et al., 2006; Weiner & Grill-Spector, 2013). In contrast, we failed to observe cross-format decoding in several functionally defined regions known to be responsive to action-relevant stimuli: verb-selective left pMTG (Bedny et al., 2008, 2012, 2014; Peelen et al., 2012; C. Watson et al., 2013) and the biological motion-selective region of pSTS (Deen, Koldewyn, Kanwisher, & Saxe, 2015; Gao, Scholl, & McCarthy, 2012; E. D. Grossman & Blake, 2002; Peuskens, Vanrie, Verfaillie, & Orban, 2005; pSTS-bio; Vaina et al., 2001). Although this latter set of null results should be interpreted with caution, it suggests that these regions might be involved in processing the lexical semantics of actions (pMTG-verb) or the motion of animate entities (pSTS-bio) rather than being involved in recognition of visual action categories per se.

Our results also accord with work suggesting that IPL is involved in abstract coding of actions. IPL has been implicated in the representation of dynamic upper-limb actions (Abdollahi, Jastorff, & Orban, 2013; e.g., Bach, Peelen, & Tipper, 2010; Cattaneo, Sandrini, & Schwarzbach, 2010; Stefania Ferri, Rizzolatti, & Orban, 2015) and tool-related actions (Mahon et al., 2007; Peeters et al., 2009, 2013; Gallivan et al., 2011; Tarhan et al., 2015; for review, see Orban and Caruana, 2014; Gallivan and Culham, 2015). Other work suggests that IPL, particularly in the left hemisphere, may represent the abstract causal outcomes or relationships between entities. For example, Oosterhof et al. (2012b) found cross-modal action-specific codes across execution and mental imagery in left IPL, but not in premotor cortex or LOTC. Left IPL exhibits adaptation when viewing reaching actions toward the same goal object, even when the hand follows a very different spatial trajectory (Hamilton & Grafton, 2006). Moreover, activation patterns in left IPL have been found to distinguish between motor acts (e.g. *pushing, grasping*) but generalize across acts performed with different body parts (Jastorff et al., 2010a), and recent work from Leshinskaya and Caramazza (2015) suggests that a dorsal portion of left IPL represents common outcomes associated with different objects even

when those outcomes are defined at a highly abstract level (e.g., a wind chime for decorating a house and perfume for decorating oneself). In our study, the spatial extent of cross-decoding was greater in the left hemisphere than the right (Table 3.1 and Figure 3.2). Taken together, previous work and the current study suggest a role for IPL (particularly on the left) in representation of actions at an abstract level.

We also observed action decoding in premotor cortex. Like LOTC and IPL, premotor cortex has been consistently implicated in action observation (e.g., Buccino et al., 2004; Gazzola et al., 2007; Saygin, 2007; Etzel et al., 2008; Majdandžić et al., 2009; Ogawa and Inui, 2011 for a meta-analysis, see Caspers et al., 2010), a finding that has been taken to support motor theories of action understanding (e.g., Rizzolatti & Craighero, 2004; Rizzolatti & Sinigaglia, 2010). In contrast, cognitive theories maintain that action understanding is achieved via higher-level, amodal representations (Caramazza et al., 2014; e.g., Hickok, 2009). Since our study only examines action observation, not action execution, we cannot address the cross-modal (observe/execute) aspects of this debate (Caspers et al., 2010; Chong et al., 2008; Dinstein et al., 2008; Kilner et al., 2009; Oosterhof et al., 2012a, 2010; Tarhan et al., 2015). Nevertheless, we did find that along with LOTC and IPL, representations of observed actions in premotor cortex were invariant to incidental perceptual features and the dynamicity of visual input. Although this results might seem superficially at odds with Wurm and Lingnau's (2015) finding that representations of *open* and *close* abstracted across the acted-upon object and the associated action kinematics in IPL and LOTC but not in premotor cortex, we believe that our result is not necessarily inconsistent. Whereas Wurm and Lingnau (2015) defined their actions by object state changes (*open* vs. *close*), we defined our actions by the physical manner of interaction (e.g., *kick* vs. *massage*). These components are logically dissociable (e.g., one can kick a door open or closed). Thus, AON regions may differ in which components of actions they represent, with premotor coding for the physical manner of action but not state-change, and LOTC and IPL coding for both. In any case, our results support the idea that there is abstraction across some features of perceptual input in all AON regions, including premotor cortex.

An open question is how the AON can extract common action codes from both static and dynamic displays. Given that in naturalistic action observation, all body parts of actors are generally visible, simple presence/absence of specific effectors in the visual

field cannot be sufficient for recognition. Instead, we hypothesize that the spatial configuration of entities (actor/effector and acted-upon entity) is crucial for determining the action category, and that parts of the AON process this configural information. Such information would be observable in both images and videos. Supporting this view, there is behavioral and neuroimaging evidence that the visual system codes the elements of actions as a perceptual unit, possibly including information about their spatial configuration, rather than simply coding them as separate, distinct items. First, briefly observed snapshots of actions are sufficient for recognition, but only when the configuration of scene entities is consistent with the given action (Dobel, Gumnior, Bölte, & Zwitserlood, 2007; Hafri et al., 2013). Second, multivoxel patterns in LOTC elicited by images of interacting humans and objects are not linearly decodable from the patterns elicited by the same actors and objects shown in isolation, yet such linear decoding is successful if the actor and objects are superimposed in a non-interacting manner (Baldassano, Beck, & Fei-Fei, 2016). This suggests that neural representations of human-object interactions (at least in LOTC) may incorporate configuration information that makes them more than the sum of their visual parts.

Another possible explanation for common static/dynamic action codes, not mutually exclusive to the above, is that through experience, static snapshots of actions become associated with full action sequences and thus elicit those sequences (Giese & Poggio, 2003; Jastorff, Kourtzi, & Giese, 2009; Singer & Sheinberg, 2010; Vangeneugden et al., 2011). This association may account for the implicit/implied motion effects observed in both behavioral and neuroimaging studies (Freyd, 1983; Gervais et al., 2010; Kourtzi & Kanwisher, 2000; Senior et al., 2000; Shiffrar & Freyd, 1993; Winawer, Huk, & Boroditsky, 2008, 2010), and may be what allows the action recognition system to be robust to missing or ambiguous perceptual input. Supporting this idea, behavioral work has shown that causal representations are engaged for both simple and naturalistic launching events despite temporary occlusion or absence of the causal moment from the stimulus display (Strickland & Keil, 2011; Yeul Bae & Flombaum, 2011).

To summarize, we uncovered abstract neural codes for action categories in bilateral OTC and IPL, left premotor cortex, and left mFG, including regions of LOTC that have been previously implicated in body-, object-, and motion- processing. These codes were invariant to differences in actors, objects, scene context, or viewpoint, and could be

evoked by both dynamic and static stimuli. Moreover, most of these regions showed consistent representational spaces across subjects and formats, which is a feature of an action recognition system that can facilitate a common understanding of actions across individuals. Taken together, our findings suggest that these regions mediate abstract representations of actions that may provide a link between visual systems that support perceptual recognition of actions and conceptual systems that support flexible, complex thought about physically interacting entities.

IV. EVENT-STRUCTURE SEMANTICS PREDICT CORTICAL RESPONSES TO NATURALISTIC LANGUAGE

1. Introduction

How the brain supports human understanding of language is a core goal of the neurobiological study of semantics. To answer this question, researchers must develop theories of how semantic information may be organized and represented, build models implementing these theories, and identify such representations in the brain by predicting brain activity elicited by semantic content. A wealth of previous work has identified areas of the brain that are candidates for semantic representation in frontal, parietal and temporal cortex, as these areas show responses selective for high-level language processing (Binder, Desai, Graves, & Conant, 2009; Fedorenko et al., 2016; Kocagoncu, Clarke, Devereux, & Tyler, 2017). Additionally, recently researchers have also made remarkable progress towards testing several classes of semantic models in the brain, suggesting that how these models represent the meanings of words may reflect or approximate how semantic information is represented in the human brain. These include "distributional semantic" (DS) models that characterize word meanings as vectors embedded in a low-dimensional space based on co-occurrence statistics (Fyshe, Talukdar, Murphy, & Mitchell, 2014; Huth et al., 2016; Mikolov, Chen, Corrado, & Dean, 2013; Wehbe et al., 2014), as well as models that characterize word meanings as a collection of sensorimotor and cognitive attributes (Anderson et al., 2017, 2018; Binder et al., 2016). However, thus far approaches in modeling brain responses to natural language have made little contact with theories in linguistics and lexical semantics. Without doing so, there is risk that a full account of human semantic organization will not be fully understood.

Lexical semantic theories address the precise ways that the structure of a linguistic utterance – how and where its parts go together – predicts its meaning, also known as semantic compositionality (Gleitman, 1990; Goldberg, 1999; Levin & Rappaport-Hovav, 2005; Pinker, 1989; Williams, 2015). The problems they address are crucial for a full, interpretable understanding of human semantic organization: After all, we do not have single isolated thoughts, such as *drop, Jimmy, ice cream cone*. We know that it is Jimmy who dropped his cone (not the other way around), and this is why he is so sad – and we

can understand as such if recounted to us by an upset Jimmy. Although there has been much work attempting to identify anatomical loci of linguistic and combinatorial operations in their own right, e.g. syntactic processing (Blank, Balewski, Mahowald, & Fedorenko, 2016; J. Brennan et al., 2012; Fedorenko, Nieto-castañón, & Kanwisher, 2013), there has been minimal work using these theories to predict how the brain responds to semantic information in a wide range of naturally occurring sentences.

In the current study, we test the hypothesis that the human linguistic system constructs a semantic representation of a verb (in part) by the semantic structure referred to by that verb. Semantic structure specifies the way that entities relate to each other and change in space and time (Jackendoff, 1990; Pinker, 1989). The semantic structure of *Jimmy dropped ice cream* can loosely be described as an event in which an Agent (Jimmy) caused the motion of a Theme (the ice cream), where the elements of this event's semantic structure are underlined. If event structure is a central component of how we represent the meaning of verbs, then event structure should be measurable in the brain while participants listen to naturalistic speech.

To test this hypothesis, we leverage previous observations in the literature that there is a strong correspondence between semantic structure and linguistic structure.[11] Linguistic structure or the syntactic frame of a verb specifies where nouns, prepositions, and other elements are situated with respect to the verb: in *Jimmy dropped ice cream*, one noun phrase (*Jimmy*) appears before the verb and one (*ice cream*) after. We test several predictions, detailed below, of what we call the Semantic Structure Consistency Hypothesis: that the semantic structure of an event referred to by a verb constrains the set of syntactic frames that a verb can (and cannot) appear in (Gleitman, 1990; Kipper, Korhonen, Ryant, & Palmer, 2008; Levin, 1993; Pinker, 1989). For example, in the sentences *Jimmy cried, Jimmy pounded the table, The father gave Jimmy another ice cream*, the number of noun phrases before/after the verb (underlined) is dictated by the nature of the events referred to: crying requires one entity, pounding two, and giving three. Furthermore, the elements that correspond between linguistic and semantic structure appear to generalize over a remarkably wide range of content differences

[11] Although other grammatical categories like nouns can refer to events (e.g. *party*) and even in some cases have semantic structure of their own (e.g., *destruction of the city*), verbs highlight the *linguistic structure / semantic structure* correspondence.

(Fisher et al., 1991; Jackendoff, 1990; Talmy, 2000). For example, although *The sun melted the ice cream* and *The Death Star vaporized the planet* vary widely in content, they both involve a common (and general) meaning: a caused change of state. The implication of the semantic structure consistency hypothesis is that verbs that share the same set of syntactic frames (such as *melt* and *vaporize*) are semantically identical, in terms of their event structure.[12]

To achieve the goal of identifying event structure representations in the brain, we implement several models based on lexical semantic theory. If the Semantic Structure Consistency hypothesis is correct, it predicts that (1) a model based on semantic structure should explain a significant portion of fMRI response variance to verbs embedded in natural language; (2) most of the response variance predicted by a semantic structure model should also be predicted by a model based on sets of syntactic frames; and (3) the response variance explained by the semantic structure model should *not* be fully explained by a simpler model that represents the single syntactic frame in which a verb is embedded in any given context. The latter two predictions hold for the following reason: any one syntactic frame on its own generally does not reveal commonalities of semantic structure, because frames can be shared by a remarkably wide set of verbs (Fisher et al., 1991; Levin & Rappaport-Hovav, 2005). For example, the transitive frame (noun-verb-noun) is very common; the verb *drop* can take it, as in *Jimmy dropped the ice cream*, but so can the verb *recount*, as in *Jimmy recounted his sad story*. But the sets of frames reveal systematic differences: *Jimmy dropped that he was sad* (strange/ungrammatical) vs. *Jimmy recounted that he was sad* (perfectly fine, though still tragic). This difference turns out to correspond to the semantic structure difference of Cause and Motion (*drop*) vs. Cause and Information Transfer (*recount*). (See Figure 4.3 in Methods and Results for details of these models.)

A secondary goal of this study is to bridge the gap between lexical semantic theory and current leading models of semantics by investigating the extent to which they predict

---

[12] To be clear, although there is hypothesized to be a *correspondence* between linguistic and semantic structure, they are not considered identical. Linguistic structure has its own set of operations and transformations (Jackendoff, 2002). Although research into semantic structure is relatively new compared to that of linguistic structure, the larger goal of research in semantics is to find elements that link the semantics specified by language with semantics in non-linguistic human cognition, such as reasoning, memory, or high-level vision.

similar variance in brain response. A schematic of the three semantic models investigated is depicted in Figure 4.1, with the Semantic Structure model depicted in Figure 4.1*A*. The first of these is an implementation of the distributional semantics approach called word2vec, depicted in Figure 4.1*B* (Mandera, Keuleers, & Brysbaert, 2017; Mikolov et al., 2013). The underlying assumption of such models is that the semantic content of a word can be inferred "from the company it keeps" (from words it co-occurs with), regardless of type (e.g., noun, verb, etc.). *Ice cream* and *cone* (two nouns) are related because these words often occur nearby the same words (e.g. *vanilla*), just as *ice cream* and *eat* (noun, verb) may too. By reducing the dimensionality of such co-occurrence information gleaned from billions of words of text, words become embedded in a dense, low-dimensional space in which neighbors in the space are assumed to be semantic neighbors. The second model comes from Binder and colleagues (Binder et al., 2016) and is based on the assumption that word meanings are a collection of sensorimotor and cognitive attributes, such as vision, audition, and cognition. A schematic of this model is depicted in Figure 4.1*C*. This model has been implemented via average human ratings of words along the relevant dimensions. Both models have been remarkably successful at predicting fMRI responses to naturalistic language (Anderson et al., 2017, 2018; de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017; Huth et al., 2016; Jain & Huth, 2018). Nevertheless, it is unknown how well these models predict responses to different types (verbs vs. nouns vs. others). Indeed, for verbs in particular, distributional semantic models generally perform worse on evaluation benchmarks in NLP (e.g. analogy tasks), as compared to nouns and adjectives (Gerz, Vulić, Hill, Reichart, & Korhonen, 2016). Likewise, the dataset of Binder and colleagues is composed principally of nouns or adjectives (out of 535 words, only 62 are verbs), so verb-specific semantics are at best only partially addressed thus far in this model. Testing the word2vec and Binder models will answer the extent to which these other model types implicitly predict similar information as semantic structure does, despite their different representational format and assumptions.

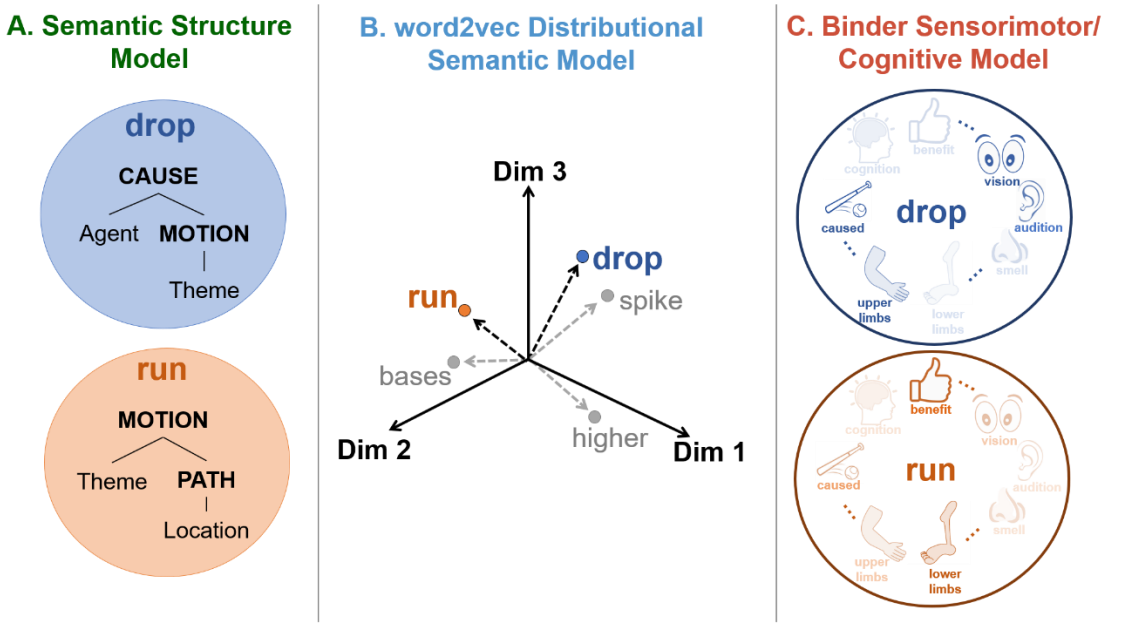


Figure 4.1
Schematic of each semantic model's representation for two example verbs: *drop* as in *Jimmy dropped his ice cream*, and *run* as in *Jimmy ran away*. **A.** The Semantic Structure model represents each verb with a set of semantic structure features such as Cause, Motion, and Path, as well as a set of event role features such as Agent, Theme, and Location. The full set of features used in this (and the other lexical semantic models used) can be found in Supplementary methods. **B.** The word2vec model represents each word as a continuous vector in a low-dimensional space in which neighbors in the space share similar linguistic contexts. Verbs are embedded in the same space as nouns, adjectives, and other grammatical types. Only three dimensions are shown here, but the dimensionality of the model used here is 300. **C.** The sensorimotor/cognitive model of Binder and colleagues (2016) represents each word as embedded in a continuous space of sensorimotor and cognitive attributes, such as vision, audition, and cognition. In the figure, the darkness of the feature image/text indicates the magnitude of that feature for the displayed word. As in the word2vec model, verbs are embedded in the same space as nouns and adjectives. Only eight features are shown here, but the full set of 23 features we used can be found in Supplementary Methods (we used a reduced set of all 65 features from the original Binder model).

The core of our approach to modeling fMRI responses to natural language is the construction and evaluation of voxelwise encoding models. Encoding models were first implemented in fMRI to study low- and high-level vision (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011) and were then extended to test models of semantics in natural language (Huth et al., 2016). Using this approach allows us to place all models "on the same footing". To objectively compare models, it is not sufficient to show that model features can be decoded in a given region, which can be possible with a high degree of accuracy even if a model explains very little variance in the fMRI response compared to others (Naselaris & Kay, 2015). Instead, we wish to directly compare how

much fMRI response variance each model explains, and the degree to which the models capture the same variance. We focus in particular on a set of language-selective regions across frontal and temporal cortex. These regions are of primary interest given they have been strongly implicated in supporting the human capacity for language (J. Brennan et al., 2012; J. R. Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010; Hickok & Poeppel, 2007).

To anticipate our results, we find significant prediction accuracy in the fMRI response to verbs in naturally occurring sentences for all lexical-semantic theoretical models, relating to both syntactic and semantic information. These effects were present throughout language-selective cortex, with the strongest results in posterior temporal cortex, consistent with prior reports of this region's responsivity to verb and event information (Bedny et al., 2008; Hafri et al., 2017; Lingnau & Downing, 2015; Peelen et al., 2012). A comparison across these models confirmed the above predictions of the Semantic Structure Consistency Hypothesis, providing support for the semantic structure approach to meaning in the brain. Additionally, we find that a majority of fMRI response variance predicted by the semantic structure model was shared with the other semantic models (word2vec and Binder). This suggests that most neural information about semantic structure is implicit in the other model representations.

2. Methods and Results

The main goal of the study was to examine the degree to which semantic structure predicts responses to natural language. We also sought to determine the information shared between the semantic structure approach and other leading approaches to semantics in the brain: word2vec (a distributional semantic model) and Binder (the sensorimotor/cognitive model of Binder and colleagues). To these ends, we recorded fMRI responses as participants listened to three hours of audiobook excerpts. We then fit encoding models in an estimation set and tested how well each predicted fMRI responses in held-out validation data. We present our results in several sections. First, we describe in detail the feature spaces used to implement the models based on lexical semantics (verb class, semantic structure, and syntactic frame models). Second, we directly compare the lexical semantic models to one another, testing the predictions of semantic structure theory. Third, we describe the other two leading models of semantics

(word2vec and Binder) and compare the representational similarity of model feature spaces to determine potential for overlap in model predictions. Finally, we directly compare the variance explained by the semantic models of interest by conducting a variance partitioning analysis. An overview of the approach and procedures used appears in Figure 4.2.

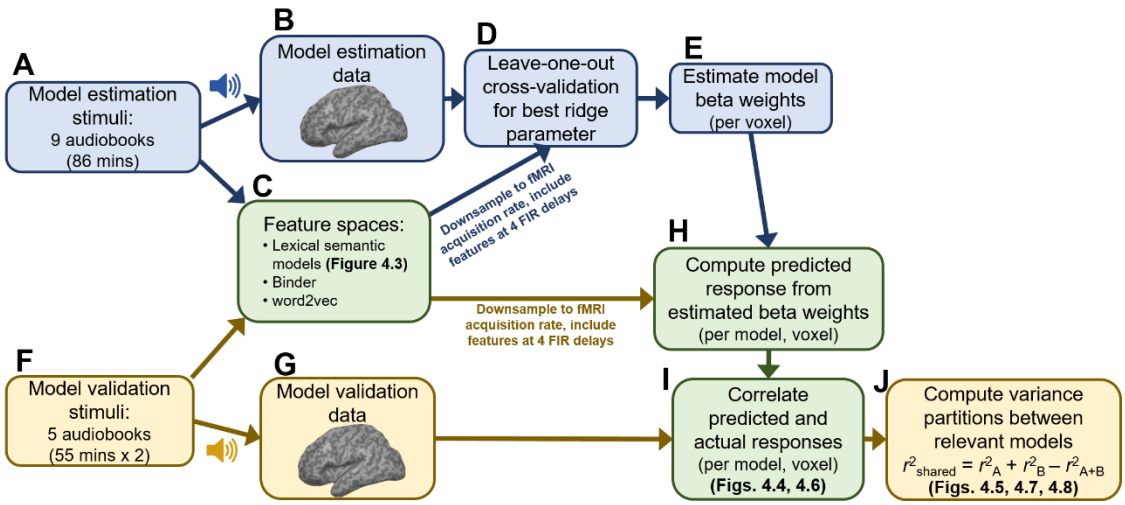


Figure 4.2

Overview of encoding model approach. The blue boxes represent steps involving model estimation data and features, the yellow boxes represent steps validation data and features, and the green boxes represent steps involving both estimation and validation data and/or features. **A.** The model estimation stimuli were 9 audiobook excerpts from different genres, each presented in a different scan run. **B.** Six subjects listened to these audiobooks as they were scanned with fMRI. **C.** The features of the stimuli were computed for each verb present in the datasets, separately for each model. Features were downsampled to the fMRI acquisition rate (2 sec per TR) and included at 4 FIR delays (2, 4, 6, and 8 seconds). **D.** Then for model fitting, the best ridge parameter was computed per voxel (for each model), by selecting the ridge parameter that yielded the highest prediction accuracy in the estimation set (via leave-one-run-out cross-validation). **E.** The full estimation set was then used to estimate the beta weights for each feature (separately for each model, and voxel). **F.** The model validation set consisted of 5 audiobook excerpts, each repeated once. **G.** The model validation data were the fMRI responses recorded as the same subjects listened to these audiobooks, averaged over the repeat scan runs. **H.** The features from the validation set (step C) and the estimated beta weights from step E were used to compute predicted voxel responses to the validation stimulus set. **I.** To assess model performance, the Pearson correlation between the predicted (step H) and actual (step G) fMRI responses were calculated (per voxel, per model). **J.** In subsequent analyses, to assess shared and unique variance attributable to each model, joint models were also fit, and variance partitioning analyses were conducted.

2.1. Description of models based on lexical semantics

To test the hypothesis that language-sensitive cortical areas encode semantic structure information present in language, we first operationalize different aspects of

linguistic and semantic structure in a set of models. To instantiate each model, we must construct feature spaces that correspond to how each model is hypothesized to represent information about linguistic stimuli. To construct these feature spaces, we utilize a database called VerbNet built by experts in linguistics (Kipper et al., 2008; Schuler, 2005). VerbNet is a database that aims to implement the hypothesis of semantic structure consistency discussed above: that there is a high correlation between the sets of syntactic frames a verb takes and its semantic structure. To this end, verbs in VerbNet are first grouped into Verb Classes based on the sets of frames each verb does or does not canonically take (Levin, 1993). For example, it is acceptable to say *Jimmy hit his father* and *Jimmy broke his ice cream cone* and *The cone broke*. It is also acceptable to say *Jimmy hit his father*, but not **Jimmy hit*. Thus, *break* and *hit* are grouped into different verb classes.

Each verb class in turn is fully described by a set of possible syntactic frames and corresponding semantic structure elements. Examples of syntactic frames might be Noun Phrase + Verb + Noun Phrase (NP V NP, or transitive), Noun Phrase + Verb (NP V, or intransitive), and Noun Phrase + Verb + Prepositional Phase (NP V PP). Examples of semantic elements are Boolean values such as Cause, Motion, and Contact, as well as event roles such as Agent and Patient. The VerbNet database allows us, for a given verb and its associated class and frame, to automatically extract feature values at these different layers. Figure 4.3 depicts how verbs and their features are assigned in VerbNet, described in more detail below.

To assign features to each verb in our audiobook stimulus set requires first identifying its specific Class and Frame based on its usage in context. Some verbs (e.g. *run*) have multiple senses, including a manner of motion (e.g. *Jimmy is running on a track*) and functioning (e.g. *the dishwasher is running*). These senses roughly correspond to different Classes. Sentential context matters even for verbs within the same class, as they can be instantiated in different syntactic frames (e.g. *Jimmy is running* vs. *Jimmy is running away from his father*). Since automated methods of identifying this information are not yet reliable (Abend, Reichart, & Rappoport, 2008; L. Chen & Eugenio, 2010; Windisch-Brown, Dligach, & Palmer, 2011), labeling each verb instance's particular Class and Frame was performed manually by the first author (A.H.). After automatic part-of-speech tagging, every verb that appeared in the VerbNet

database was manually labeled with Class and Frame. The semantic structure information (e.g., Cause, Motion, Contact) was not visible during the labeling so could not be used to bias the annotation process. (Verb-like elements such as the copula (*be*), modals (e.g. *should, must*), and auxiliary verbs (*do, have, will)* were excluded. Automatic part-of-speech tagging was verified and errors were corrected before annotation.)

			Verb Class			Single Frame					Average Frame					Semantic Structure							
			Model Features																			
Example Number	Verb (in sentence)	Frame	roll-51.3.1	say-37.7	run-51.3.2	NP pre-V	NP(1) post-V	NP(2) post-V	PP post-V	S post-V	NP pre-V	NP(1) post-V	NP(2) post-V	PP post-V	S post-V	Agent	Theme	Topic	Recip-ient	Loca-tion	Cause	Motion	Trans-fer Info
1	*Jimmy dropped the ice cream.*	NP V NP	1	0	0	1	1	0	0	0	1	0.5	0	0	0	1	1	0	0	0	1	1	0
2	*The ice cream dropped.*	NP V	1	0	0	1	0	0	0	0	1	0.5	0	0	0	0	1	0	0	0	0	1	0
	…																						
3	*Jimmy recounted the story.*	NP V NP	0	1	0	1	1	0	0	0	1	0.25	0.25	0.25	0.75	1	0	1	0	0	1	0	1
4	*Jimmy recounted to his friend that he was sad.*	NP V PP S	0	1	0	1	0	0	1	1	1	0.25	0.25	0.25	0.75	1	0	1	1	0	1	0	1
	…																						
5	*Jimmy ran.*	NP V	0	0	1	1	0	0	0	0	1	0	0	0.25	0	0	1	0	0	0	0	1	0
6	*Jimmy walked.*	NP V	0	0	1	1	0	0	0	0	1	0	0	0.25	0	0	1	0	0	0	0	1	0
7	*Jimmy walked to the store.*	NP V PP	0	0	1	1	0	0	1	0	1	0	0	0.25	0	0	1	0	0	1	0	1	0

1 Feature value 0

Figure 4.3

Table illustrating how different verbs in sentential contexts are coded in each of the models based on lexical semantic theory. In the example sentences, the verb of interest is bolded, and each segment of underlined text indicate a phrase (e.g., the ice cream in example 1 is a Noun Phrase). Each verb or phrase in the example sentence corresponds to a syntactic element in the Frame column. In example 4, that he was sad is a sentence complement (subordinate clause) of the verb recount; the verb of the sentence complement is not coded here. Notice several properties of the feature coding. First, the Verb Class and Average Frame models are constant for each verb no matter the specific frame the verb is embedded in. In contrast, both the Single Frame and Semantic Structure model coding of each verb example vary slightly depending on sentential context. Example 2 (with the verb drop) has similar coding to example 1 but it has no NP post-verb feature in the Single Frame model, and no Agent or Cause features in the Semantic Structure model. Such differences can also be observed in examples 3 and 4, and 5-7. Also note that the presence of the same syntactic frame element (e.g. PP post-verb) corresponds to different features in the Semantic Structure model, depending on the Verb Class: in example 4, it corresponds with Recipient, while in example 7, it corresponds with Location. Notice also that different verbs can have identical class, frame, and semantic structure features, as in examples 5 and 6 (run and walk): as far as these models are concerned, these examples have identical representations. Finally, notice that despite some semantic structure features being present or absent depending on the sentential context, some remain as "core" semantic structure features for that class. For example, for examples 1 and 2, as well as 5-7, it is Motion; for examples 3 and 4, it is Transfer_Information. Together these examples illustrate that there is a strong relationship between Average Frames and Semantic Structure, although the mapping between Syntactic Frame features and Semantic Predicate features is many-to-many. Abbreviations: NP, Noun Phrase; PP, Prepositional Phrase; V, verb; S, sentence complement.

The first linguistic model is a Verb Class model. This model instantiates the groupings of verbs based on their shared sets of syntactic frames, but without additional information about the precise content of those syntactic frames or the associated semantic structure predicates. In other words, verbs that are grouped into the same class

based on their shared frames, such as the verbs *run, walk, limp*, will each have an identical Class feature assigned (in VerbNet, *run-51.3.2*). However, information about their frames (e.g. NP V, NP V PP) and semantic predicates (Motion) – information that cuts across verb classes – are not explicit features in this model. Given the lack of latent structure between classes, this model is not expected to do as well as the other models below. Examples of Verb Class assignment appear in Figure 4.3.

The second model is a Single Frame model. This model instantiates the surface syntactic frame of a verb as it appears in a particular sentential context.[13] For example, in a sentence such as *Jimmy dropped the ice cream*, the frame coding is NP V NP. In contrast, in a sentence such as *The ice cream dropped*, the frame coding is NP V. Examples of feature assignment for this model appear in Figure 4.3. Note, as was delineated previously, that it is hypothesized that the *set* of syntactic frames a verb takes that corresponds to its underlying semantic structure, rather than any single instance. Thus, this model is hypothesized not to perform as well as a model based on the *sets* of syntactic frames a verb takes – the Average Frame model, described below.

The third model is an Average Frame model, which is also a syntactic model. The features in this model are the average of the set of surface syntactic frames for a verb given its class. For example, at an instance of the verb *drop*, its class will be looked up, and an average over all frames for that class computed (in this case, NP V, NP VP NP, NP V PP, among others). Although in principle certain frames may be more important for the representation of a verb than others (and may be based on the frequency of such frames), here we make a simplifying assumption that they all receive equal weight. Examples of feature assignment for this model appear in Figure 4.3.

The exact features in both syntactic models (Single and Average) were the syntactic *elements* in the frame, dependent on its position in relation to the verb and how many of that type appeared pre- vs. post-verb. This can be considered a proxy for its syntactic structure. In the case of NP V NP, the verb would get the features NP_pre-verb_1,

---

[13] By surface syntax, we mean the "core elements" of a verb's frame, apart from syntactic transformations (like the passive) and additional non-essential elements (i.e. adjuncts). For example, the surface structure of a frame (e.g. noun-verb-noun) may be transformed by syntactic operations: *Jimmy dropped ice cream* can be transformed to a passive expression, *Ice cream was dropped by Jimmy*. Additionally, there is a lot of information that can be included in an utterance that is non-essential to the interpretation of the main verb, e.g. the underlined text in *Jimmy, <u>who cries all the time</u>, dropped his melting ice cream*.

NP_post-verb_1; for NP V NP NP, it would get the same features plus NP_post-verb_2; NP V NP PP would get a PP_post-verb_1 feature. Implementing the model in this way captures any shared information across frames that could be due to the elements of the syntactic frame itself (for example, a PP post-verb may indicate a change of location across multiple verb classes). Otherwise, if we just use the full frame without decomposing it into elements (e.g. "NP V NP" could be considered Frame 1, "NP V NP PP" Frame 2, with no relation between them), this might fail to capture shared elements of meaning contributed by particular frame elements. See Figure 4.3 for examples of how feature assignment proceeds in these models.

Finally, the fourth model is a Semantic Structure model. The experts who created the VerbNet database examined the sets of verbs in each class and determined from these the semantic structure elements for each, based on commonalities of meaning from previous lexical semantic literature. For example, verbs like *melt* or *redden* involve a change of State; verbs like *hit* and *touch* involve Contact. Note here that the elements at this layer will only be as good as the labels provided by the expert annotators; some elements (like Motion, State, or Contact) are more clearly identifiable and easily labeled than others (Fisher et al., 1991; Hartshorne, Bonial, & Palmer, 2014). Despite this drawback, we can still extract the elements at this layer. This model instantiates semantic structure in the following way. Based on the Class of a given verb in context, the semantic predicates for that class and frame are extracted. For example, in *Jimmy dropped the ice cream*, predicates Cause, Motion, and semantic roles Agent (the one acting) and Theme (the one undergoing change/motion) are extracted. In *The ice cream dropped*, only the Motion and Theme elements are explicit. As can be observed, the full set of semantic elements will vary depending on the frame, even within class; however, some elements will nearly always be present (such as Motion and Theme here). Examples of feature assignment for this model appear in Figure 4.3.

Before continuing, it is worth briefly expounding on the relationship between syntactic frame elements and semantic structure elements: although related (Goldberg, 1999), they are in a many-to-many relationship (i.e. they are not exact supersets or subsets of one another). First, the *same* frame elements can correspond to *different* semantic elements, depending on the verb class. For example, verbs that can take identical syntactic frames (e.g. NP V NP) such as *Jimmy dropped the ice cream* and

Jimmy recounted the story, will have a different set of semantic elements depending on their class: for *drop*, Cause, Motion, Agent, and Theme; for *recount*, Cause, Transfer_Information, Agent, and Topic. Likewise, different frame elements can correspond to the same semantic elements. For example, in *Jimmy dropped the ice cream* (NP V NP syntactic frame), the post-verb NP is the Theme; but in *The ice cream dropped* (NP V frame), the pre-verb NP is the Theme. Thus, although frame elements and their positions are correlated with semantic predicates (e.g., a Theme will most often be a pre- or post-verb NP, not a PP), they capture divergent sets of information. See Figure 4.3 for a visualization of this.

2.2. Evaluation and comparison of the models based on lexical semantic theory

After creation of the feature spaces, we used ridge regression to fit all models to the estimation data set (9 of the 14 audiobook scan runs), separately for each voxel. The best ridge parameter was selected per voxel per subject via leave-one-run-out cross-validation across scan runs. Subsequently, we used the fit models to predict the fMRI response in each voxel for the held-out validation set (5 audiobook runs). We then compared the predicted fMRI response to the actual fMRI response in each voxel (the 5 runs, averaged across 2 repetitions) by calculating the Pearson correlation (r) between the predicted and actual responses in the validation data. The mean correlation across voxels in each ROI for each model was the measure of interest.

We focused our analyses on ROIs defined in an independent functional localizer contrasting intact spoken language with closely matched uninterpretable speech (Fedorenko et al., 2010; Scott, Gallée, & Fedorenko, 2016). This localizer identifies a set of language-selective regions in prefrontal, temporal, and parietal cortices: inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), middle frontal gyrus (MFG), anterior temporal (AntTemp), middle anterior temporal (MidAntTemp), middle posterior temporal (MidPostTemp), posterior temporal (PostTemp), and angular gyrus (AngG). We evaluated each model for its consistency with predictions made by lexical semantic theory, as described below.

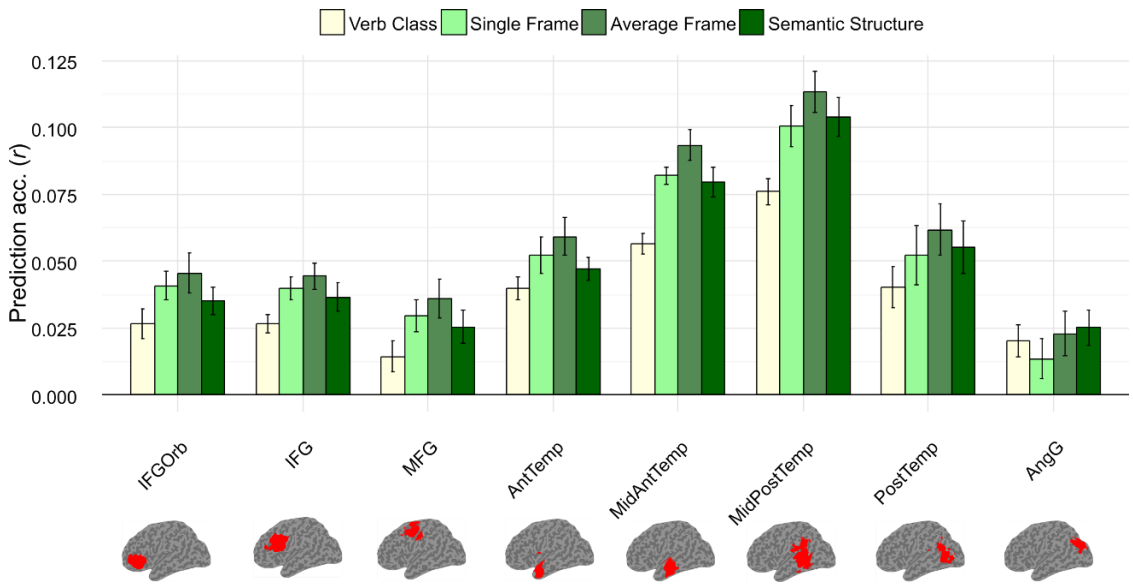


Figure 4.4

Prediction accuracy by ROI for the four models based on lexical semantic theory. Bars show the mean Pearson *r* value across subjects for each ROI. Brain maps below ROI labels indicate the parcels used to make subject-specific language-selective ROIs, based on selectivity from an independent functional localizer. Error bars are the standard error of the mean across subjects. The Average Frame model showed greater prediction accuracy than the Single Frame model in all ROIs. Both the Average Frame and Semantic Structure models showed greater prediction accuracy than the Verb Class model in all ROIs except AngG (and for Semantic Structure model, AntTemp and PostTemp as well). The Average Frame model showed greater prediction accuracy than the Semantic Structure model across ROIs. Abbreviations: inferior frontal gyrus orbital (IFGOrb), inferior frontal gyrus (IFG), middle frontal gyrus (MFG), anterior temporal (AntTemp), middle anterior temporal (MidAntTemp), middle posterior temporal (MidPostTemp), posterior temporal (PostTemp), and angular gyrus (AngG).

Figure 4.4 shows average prediction accuracy across subjects, for each ROI and each model. All models make significantly accurate predictions in all ROIs (p's < .05, confirmed by repeated-measures ANOVA with ROI as a factor, followed by post-hoc paired t-tests in each ROI), except AngG for both the Single Frame model (p = .11) and Average Frame model (marginal at p = .06). The middle anterior and posterior temporal cortex ROIs (MidAntTemp and MidPostTemp) showed especially strong predictions. This is expected, as this set of regions has previously been observed to show robust encoding of semantic information elicited by linguistic input (Bedny et al., 2008, 2014; Binder et al., 2009; Huth et al., 2016). These results thus far suggest that information related to lexical semantics is encoded in these regions.

To understand the precise information content of these regions in relation to lexical semantics, we conducted several additional analyses. First, if the set of syntactic frames

is what is crucial for predicting semantic information elicited by verbs in the brain, then we should find that the Average Frame model performs better than the Single Frame model. This was indeed the case. In all ROIs, the Average Frame model showed higher prediction accuracy than the Single Frame model, nearly significant at the .05 level (repeated-measures ANOVA with ROI and Model Type as factors, main effect of Model Type: $F(1,5) = 6.12$, $p = 0.056$), as can be observed in Figure 4.4. This may be somewhat surprising, given that we might expect such language-sensitive regions to be sensitive to the particular syntactic context in which a verb is embedded a sentence (for evidence that these regions show differential responses to syntactic manipulations, see J. Brennan et al., 2012; J. R. Brennan et al., 2016). Instead, at least as can be observed in the current data, this result suggests that the *set* of frames permissible for a verb is key for predicting responses to verbs compared to a model based on the particular syntactic frame of a verb in context.

Second, if the shared semantic or syntactic elements of verb classes is what is predictive of fMRI response, rather than the mere grouping per se, then the models with semantic and syntactic elements – the Semantic Structure and Average Frame models – should show higher prediction accuracy than a model based only on Verb Class assignments, which contains no information about shared elements across classes, whether linguistic or semantic. This result was confirmed. As can be observed in Figure 4.4, the Average Frame model showed higher prediction accuracy than the Verb Class model, in the majority of ROIs (confirmed by a repeated-measures ANOVA, interaction of ROI and Model Type: $F(7,35) = 7.35$, $p < .001$). Post-hoc paired *t*-tests in each ROI showed that all ROIs showed such an effect (p's $< .05$) except AngG ($p = .62$). Similar results were observed for the Semantic Structure model (interaction of ROI and Model Type: $F(7,35) = 5.51$, $p < .001$). Post-hoc paired *t*-tests in each ROI showed that the majority of ROIs showed such an effect (p's $< .05$), except for AntTemp ($p = .07$), PostTemp ($p = .10$), and AngG ($p = .25$). Thus, the grouping of verbs into classes based on shared syntactic and semantic structure does not alone predict fMRI responses at the level of models incorporating the content (whether syntactic or semantic) of what membership in a class entails.

The next question we sought to address is the relationship between syntactic frame models and semantic structure model. The predictions of the Semantic Structure

Consistency Hypothesis are that the sets of syntactic frames a verb takes together predict semantic structure of the verb, not just the single frame of the verb in context. In other words, despite the fact that the models traffic in qualitatively different feature types (elements of syntactic frames, e.g. NP, PP, vs. elements of syntactic structure, e.g., Agent, Theme, Cause), they should explain similar (or the same) variance. This is not a given; due to their different feature types, it is certainly possible that the two model types explain different variance in the fMRI response. Perhaps the frame models are actually explaining only syntactic information, while the semantic structure model is explaining semantic information, and thus they are explaining two different aspects of the fMRI response.

To formalize whether the Frame and Semantic Structure models are explaining the same or different variance of the fMRI response, we conducted two variance partitioning analyses within each ROI. Variance partitioning allows one to determine whether sets of models predict unique or shared variance by attributing variance to each model according to whether a joint model leads to a gain in variance explained or not (de Heer et al., 2017; Lescroart & Gallant, 2018). The logic is, if one or both models explain independent variance in the fMRI response, then when their feature spaces are combined together in the joint model, this model should show a gain in explained variance. In contrast, if they explain the same variance, then the joint model should show no gain in variance explained. We performed a variance partitioning analysis for Single Frame and Semantic Structure models, as well as Average Frame and Semantic Structure models.

The variance partitioning analyses revealed several insights. First, for both variance partitioning analyses (Single Frame with Semantic Structure, and Average Frame with Semantic Structure), a majority of the variance predicted by either frame model and the Semantic Structure model was shared, as can be observed in Figure 4.5. Second, and crucially, there was a tradeoff between unique variance for the Semantic Structure model and its shared variance with two frame models: when average frames were used instead of single frames, some of the unique variance attributed to the Semantic Structure model was “soaked up” by both the shared variance between the frame and semantic structure models, and the frame model itself. This was confirmed in a repeated-measures ANOVA with the factors ROI, frame model type (Single vs. Average), and the particular variance

partition (Shared, Unique to frame model, or Unique to Semantic Structure model): the ANOVA revealed an interaction between the model type and variance partition ($F(2,10) = 5.66$, $p = .02$) with a marginal triple interaction with ROI ($F(14,70) = 1.71$, $p = .074$). Three post-hoc ANOVAs comparing the pairwise variance partitions (collapsing across ROIs) revealed that the differences were attributable to an increase in both the shared variance and unique variance of the frame model, along with reduced unique variance for the Semantic Structure model: the interaction of frame model type and partition was significant when comparing Shared variance and Semantic Structure unique variance ($F(1,5) = 6.56$, $p = .05$) and Frame model and Semantic structure unique variances ($F(1,5) = 6.27$, $p = .05$) but not Shared variance and Frame model unique variance ($F(1,5) = 0.77$, $p = .42$). In other words, there was more in common between a Semantic Structure model and a syntactic frame model once sets of frames were taken into account (i.e. the Average Frame model) rather than the particular frame in context (i.e. the Single Frame model). This confirms the predictions of the Semantic Structure Consistency Hypothesis.

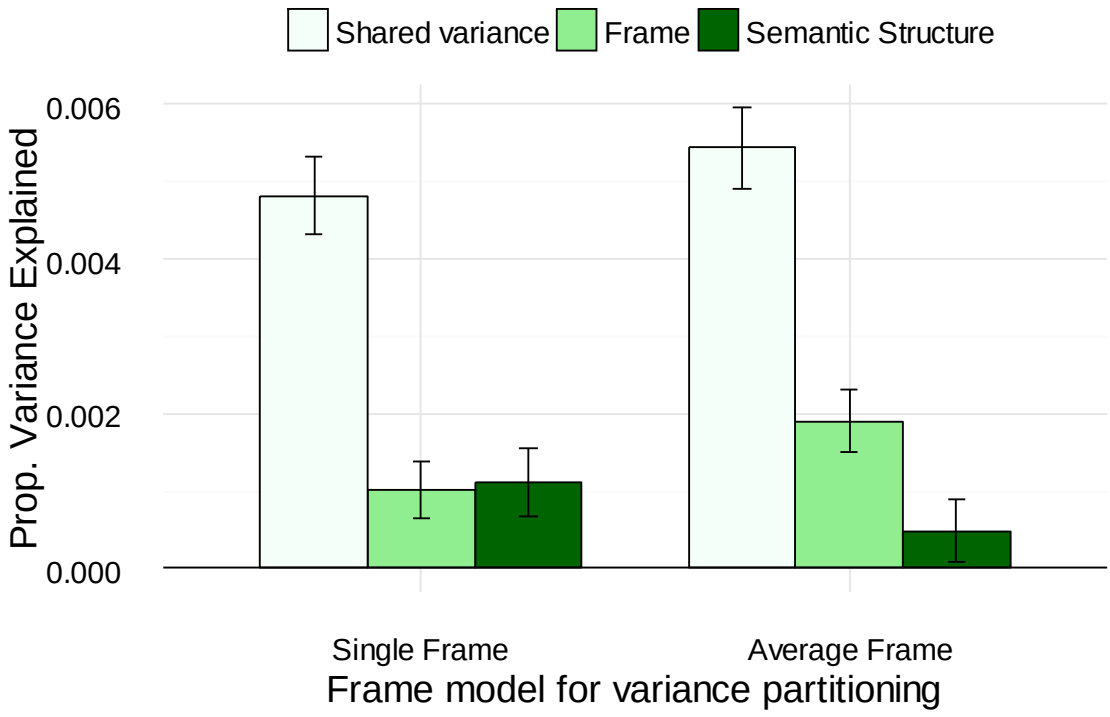


Figure 4.5
Variance partitioning analyses between the Semantic Structure model and two different syntactic frame

models. One variance partitioning analysis was with the Single Frame model (left) and the other variance partitioning analysis was with the Average Frame model (right). Joint models between pairs of models were run, and variance was partitioned using set theory to each single model (unique variance) or pair of models (shared variance), accordingly. Bars show the mean proportion of variance explained across subjects for each variance partition. Error bars are the standard error of the mean across subjects. Results are averaged across ROIs because there was only a marginal interaction of partition and frame model type with ROI (p = .074) and interaction between partition and frame type is more easily observable averaged across ROIs. This interaction can be observed as the trade-off between the unique variance attributable to the Semantic Structure model (dark green bars) on the one hand, and the Frame model (light green bars) and Shared variance (white bars), on the other, dependent on whether the frame model is the Single (left) vs. Average (right). The Average Frame model absorb variance initially attributable to the Semantic Structure model (in its analysis with the Single Frame model): the unique variance for the Semantic Structure model decreased while the shared variance and unique variance attributable to the frame model increased when the Average rather than Single frames were used.

One final aspect of the data needs to be addressed here. We find that the Average Frame model performed better than the Semantic Structure model across all ROIs, as can be observed in Figure 4.4. This was confirmed in a repeated measures ANOVA as a main effect of Model Type, $F(1,5) = 6.16$, $p = .056$. (There was a trend towards an interaction of ROI and Model Type as well, $F(7,35) = 2.10$, $p = .07$). Thus, the model with *sets* of syntactic frames predicts fMRI responses better than a modeling incorporating elements of semantic structure. We suspect that this may relate to deficiencies with identifying and labeling the elements of semantic structure: such a process can be challenging even for experts, as discussed further in the General Discussion (Fisher et al., 1991; Hartshorne et al., 2014). In contrast, labeling the elements of syntactic structure is trivial and can be automated; it simply involves labeling words in a frame with their syntactic class (noun phrase, prepositional phrase, etc.). Although the Average Frame model performs significantly better than the Semantic Structure model, both models show significant prediction accuracy, and the Semantic Structure model has the advantage of interpretability of its semantic content. Thus, in the next section, we use the Semantic Structure model rather than the Average Frame model for further comparison to other leading semantic models. Results are qualitatively similar whether we use the Semantic Structure model, or the Average Frame model as a proxy for semantic structure. We return to these issues in the General Discussion.

2.3. Description of feature spaces for other semantic models

Our analyses thus far support the conclusion that semantic structure is a valid characterization of the representations elicited by verbs in naturalistic linguistic input.

We next sought to characterize the semantic structure model's relation to two other leading semantic models (depicted in Figure 4.1). In this section, we describe the feature spaces of the comparison models, we relate the feature spaces across the three models, we test the prediction accuracy of each model, and we determine the degree to which each model explains unique variance in the fMRI response to linguistic input.

The first model is a distributional semantic model (*word2vec*, 300 features; Mikolov et al., 2013). The underlying assumption of this and related models is that the lexical neighborhood of a word reveals the semantics of the word. For word2vec, dense vectors for each word are learned by a neural network trained to predict the neighboring words of a given target word in a large corpus of over 100 billion words of text. After training, each word is represented as a vector in a relatively low-dimensional continuous space, where neighboring words in the space are assumed to share a semantic representation. Each feature is constrained to have a value between -1 and 1. This and similar distributional models have been found to predict human performance on semantic tasks (Mandera et al., 2017), and to predict fMRI responses to naturalistic speech (Huth et al., 2016; Jain & Huth, 2018; Pereira et al., 2018). Details on this model have been described elsewhere and so will not be described further here.

The second model is a theoretical semantic model based on sensorimotor and cognitive domains known to be relevant to human cognition (e.g., motion, sensation, audition, communication), developed by Binder and colleagues (2016). The model was operationalized based on average human ratings for individual words (434 nouns, 62 verbs, 39 adjectives) along 65 features of interest; for example, the degree to which the word *swatted* involves seeing something (vision), on a scale from 0 to 6. This model has been previously shown to predict fMRI responses to language stimuli (Anderson et al., 2017, 2018), and it is a leading model in the field of neurobiology of semantics. In our implementation, we obtained new average ratings for our set of verbs (675) along a reduced set of 23 features, chosen based on a factor analysis in Binder et al. (2016) where such features maximized variance in the feature set. Our ratings conform well to Binder's original ratings (across 47 verbs common to Binder's ratings and the current study, ratings correlated on average $r = 0.89$). This model has already been described in detail in previous work, so will not be described further here.

2.4. Examination of feature spaces of semantic structure model and other models

Before examining the relationship between the Semantic Structure model and the other two semantic models (word2vec and Binder), we wished to quantify their potential to explain common variance in the fMRI response to our language stimuli. Since the features in each model are not directly comparable on their own, our general approach to this is RSA, an analysis technique for comparing representational spaces that differ in number and kind of feature (Kriegeskorte et al., 2008). By using an abstract representational space in which similarities of stimuli or categories are represented (instead of similarities of features themselves), models can be compared in terms of second-order similarity, without a need to assume a precise mapping between features in one space (e.g. behavior) and another (e.g. model).

Three representational dissimilarity matrices (RDMs) were constructed, in the following way. For each model, we extracted the feature channel timecourses after pre-processing (4,318 total timepoints). Each RDM was computed by constructing the pairwise similarity matrix of each timepoint to every other timepoint based on squared Euclidean distances of each observation's set of features. Thus, the RDM for each model was a timepoint x timepoint dissimilarity matrix, representing how similar or different the model "considered" each timepoint. The lower the distance, the more similarly the timepoints are represented in the model.

To determine the similarity across models, we next computed the pairwise RDM correlations across models (using the lower off-diagonal values of their RDMs). This analysis revealed that the Binder and word2vec models were correlated at $r = 0.56$, the Binder and Semantic Structure models at $r = 0.25$, and the word2vec and Semantic Structure models at $r = 0.47$. This analysis suggests the possibility of substantial overlap in variance explained between these models. Similar correlation values were found when comparing RDMs constructed from the feature values for each instance of a verb independent of its temporal placement in each scan run (3,146 verb instances total). This suggests that the observed correlations were not simply due to an artifact of downsampling the feature channels to the acquisition rate of fMRI.

Despite the similarities between models, it is important to note that finding these correlations does not necessitate that the correlated information between the model

feature spaces is the same as that which predicts fMRI responses to language. In other words, imagine that the information not correlated between the models (i.e. the unique aspects of word2vec and Semantic Structure models) is what predicts fMRI responses to language separately. In that case, these models would predict little of the same variance. Conversely, if the aspects of the models that predict fMRI activity are the same as those which result in the correlation between model RDMs, then the models have the potential to explain the same variance in fMRI response. To examine this, we turn to the performance of each model at predicting fMRI response, and then conduct a variance partitioning analysis among the three models to determine the unique vs. shared contribution of each to predicting fMRI response.

2.5. Evaluation of semantic structure model and other models

Next, we evaluate the predictive power of each model type. We fit each model to the estimation data set using ridge regression and calculated the Pearson correlation between the predicted and observed fMRI responses in the validation data set. Predictions for the Semantic Structure model are the same as in the previous section and are recapitulated in Figure 4.6, along with predictions for the other two models. The word2vec model resulted in significant prediction accuracy across all ROIs (all p's < .01), replicating previous work using word embeddings (Huth et al., 2016). Notably, this prediction accuracy was quite high despite the fact that unlike in these previous studies, only verb timepoints contributed features for prediction; all other words did not contribute features for modeling. The Binder model also showed significant prediction across ROIs (all p's < .05). A repeated-measures ANOVA comparing mean prediction accuracy across ROIs between the models found a main effect of Model Type ($F(2,10) = 8.10$, $p = .008$), with no interaction across ROIs ($F(14,70) = 0.92$, $p = .55$). Post-hoc paired t-tests between each pair of the three models showed that word2vec performance was significantly higher than both Binder ($t(5) = 3.94$, $p = .01$) and Semantic Structure ($t(5) = 3.08$, $p = .03$); there was no significant difference between Binder and Semantic Structure ($t(5) = 1.72$, $p = .15$). These results suggest that the three models provide a good description of the representation elicited by verbs in naturalistic language.

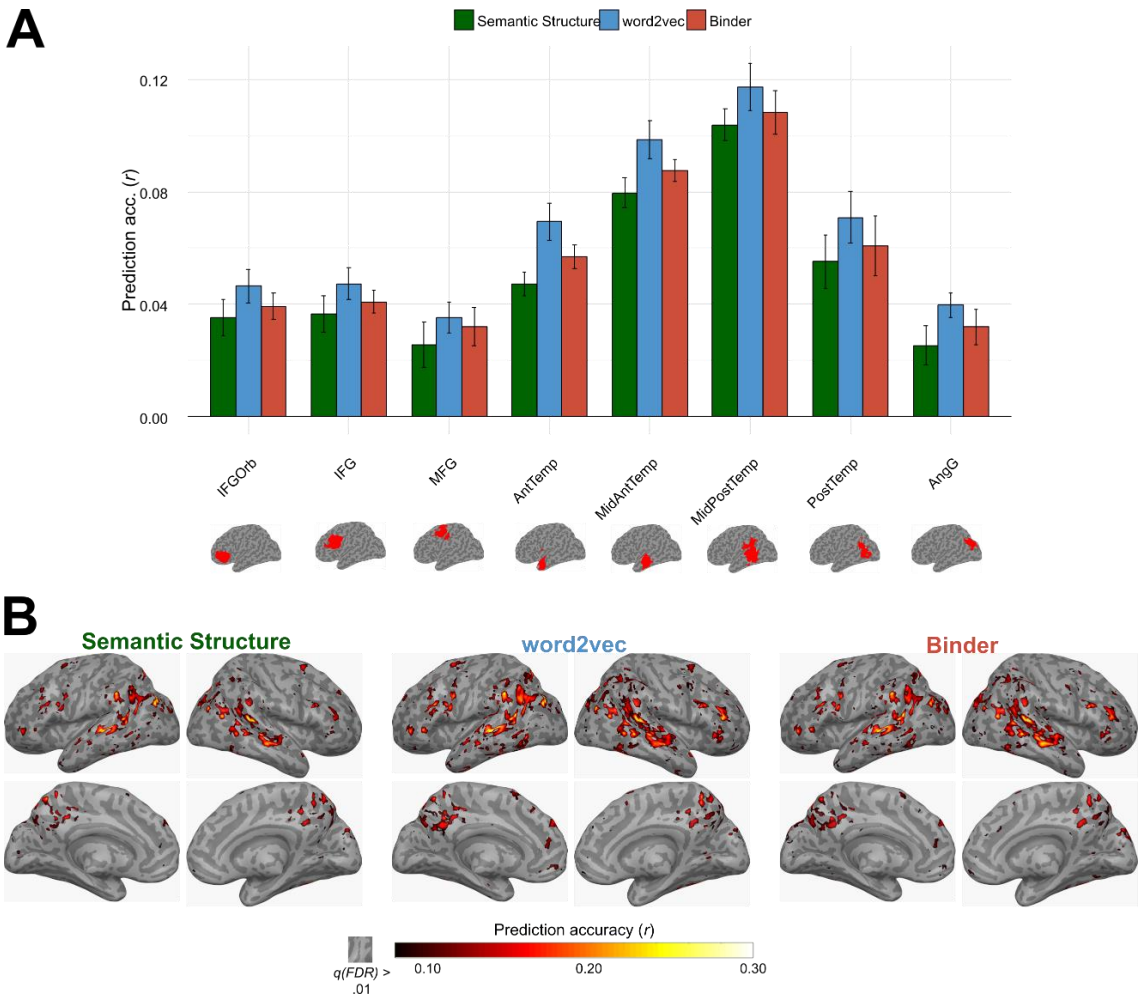


Figure 4.6

Prediction accuracy for the three semantic models of interest. **A.** Bars show the mean Pearson *r* value across subjects for each ROI. Brain maps below ROI labels indicate the parcels used to make subject-specific language-selective ROIs, based on selectivity from an independent functional localizer. Error bars are the standard error of the mean across subjects. Across all ROIs, word2vec showed significantly greater prediction accuracy than the other two models ($p < .05$). **B.** Whole-brain maps of prediction accuracy for one subject, plotted on the inflated cortical MNI surface. Prediction accuracy is the Pearson correlation between predicted and actual fMRI responses to the validation stimulus set. Significant predictions for the three models were observed in similar areas, including outside of language-selective regions, such as medial parietal areas and some ventral regions. Abbreviations: inferior frontal gyrus orbital (IFGOrb), inferior frontal gyrus (IFG), middle frontal gyrus (MFG), anterior temporal (AntTemp), middle anterior temporal (MidAntTemp), middle posterior temporal (MidPostTemp), posterior temporal (PostTemp), and angular gyrus (AngG).

2.6. Comparison of semantic structure model and other models

Even though the models each make significant predictions across ROIs (albeit at slightly different levels), prediction accuracy alone does not reveal the degree to which

the models explain independent variance. That is, it is possible that all three models quantify the same amount of variation in the linguistic input. But in different ways; for example, perhaps the Semantic Structure model quantifies variance due solely to properties of events such as Cause and State Change, while the Binder model explains variance based on representations not related to event structure, such as the particular sensory system involved (indeed Semantic Structure considers verbs of perception such as *see a*nd *hear* to be identical, while in the Binder model there is a set of features along which these differ: vision vs. audition). To address the degree to which the models explain shared or unique variance, we conducted a variance partitioning analysis, as in the previous section on lexical semantic models (de Heer et al., 2017; Lescroart & Gallant, 2018). This analysis involves fitting joint models (e.g., a model with both Semantic Structure and Binder feature sets), and then using set theory to allocate variance between models depending on the degree to which the joint model results in a gain in variance explained.

Results of the variance partitioning analysis appear in Figure 4.7. The analysis revealed several key properties of the explained variance of the models. The main finding was that the greatest variance partition was the shared variance of all three semantic models (Semantic Structure, Binder, and word2vec; pairwise paired *t*-test comparisons among all seven partitions, all *p*'s < .006 uncorrected, all *p*s < .09 Holm-Bonferroni corrected). This result suggests that these semantic models in large part capture similar aspects of the representation of verbs present in natural language. It also confirms our previous representational similarity analysis of the feature spaces of each model, where we found the feature spaces themselves were correlated across verb instances in our stimuli with *r* values between 0.25 and 0.56. We think it is likely that by the nature of its training, word2vec has likely implicitly learned information related to semantic structure. Likewise, since Binder also shares a substantial amount variance with the other two models, this suggests that when people give judgments about properties of verbs (the input to the Binder model), they have implicit access to event structure for making such judgments. We return to these issues in the General Discussion.

Second, we tested for evidence of significant explained variance by each shared and unique model partition. One important note of caution here: because variance partition estimates were computed as variance explained in the held-out validation set, and such

estimates will necessarily contain sampling noise, using the set theory approach for variance partitioning has the potential to result in mathematically impossible results, i.e. a variance partition estimate below zero. Although this results in some impossible values (e.g., the shared variance of Binder and Semantic Structure being significantly negative), adjusting for this by estimating a bias term (for each voxel), as in (de Heer et al., 2017), has its own issues, i.e. it may inflate significance of results across subjects (because no value can be below zero). Thus, we chose instead not to adjust the partition estimates. These results provide a lower bound on the possible unique variance attributable to each semantic model or sets of models. If a region shows variance explained significantly *greater* than zero, we can be confident this is indeed the case, but a null result here would be inconclusive.

To determine the significance of explained variance for each partition, we conducted individual ANOVAs for each partition comparing variance explained across ROIs, followed by post-hoc *t*-tests for each ROI if an interaction of ROI was significant; one individual *t*-test across all ROIs otherwise (this was the case for the unique variance of the Binder model, and the shared variance of Semantic Structure and word2vec models). *T*-tests were one-sided, testing whether explained variance was significantly greater than zero. First, the shared variance among all models was significant in each ROI (all *p's* < .02; white bar in Figure 4.7). Second, we found evidence that individual models, and pairs of models, capture significant explained variance not attributable to the others. Shared variance between Binder and Semantic Structure was not significantly greater than zero in any ROI (*p*'s > .86, orange bar in Figure 4.7). Shared variance between Binder and word2vec was significant in the majority of ROIs (*p*'s < .05, purple bar in Figure 4.7) except in IFG, where it was marginal ($p = .07$). There was no significant shared variance between Semantic Structure and word2vec (no difference among ROIs, test on mean across all ROIs $p = .95$, light green bar in Figure 4.7). Unique variance attributable to Semantic Structure was significant in all ROIs ($p < .05$, dark green bar in Figure 4.7) except MFG, AntTemp, and AngG (*p*s > .16). Additional unique variance attributable to word2vec was significant ($p < .05$) in all ROIs (red bar in Figure 4.7), marginal in AntTemp ($p = .08$), and not significant in MFG or AngG (*p's* > .28). Finally, Binder showed no unique variance in any ROI (no difference among ROIs, test on mean across all ROIs $p = .57$, blue bar in Figure 4.7). Whole-brain maps of the variance

partitioning analyses are shown in Figure 4.8 for one example subject.

Together, this variance partitioning analysis shows that the greatest partition of variance explained was shared variance attributable to all three models; in other words, their feature sets shared relationships (at least in our data set) such that they explained the fMRI response to verbs equally well. For additional variance attributable to individual or pairs of models, strongest results were found in the anterior/middle temporal regions (AntTemp, MidAntTemp, and MidPostTemp) and IFG regions (IFG, IFGOrb). For these, there was significant variance that could be attributable jointly to Binder and word2vec, as well as additional unique variance attributable solely to the word2vec model or the Semantic Structure model that could not be accounted for by the other models. Although we cannot conclusively rule out unique variance attributable only to the Binder model due to the bias downward of variance partitioning conducted on the validation set (discussed above), we did not find evidence for additional significant variance solely due to the Binder model.

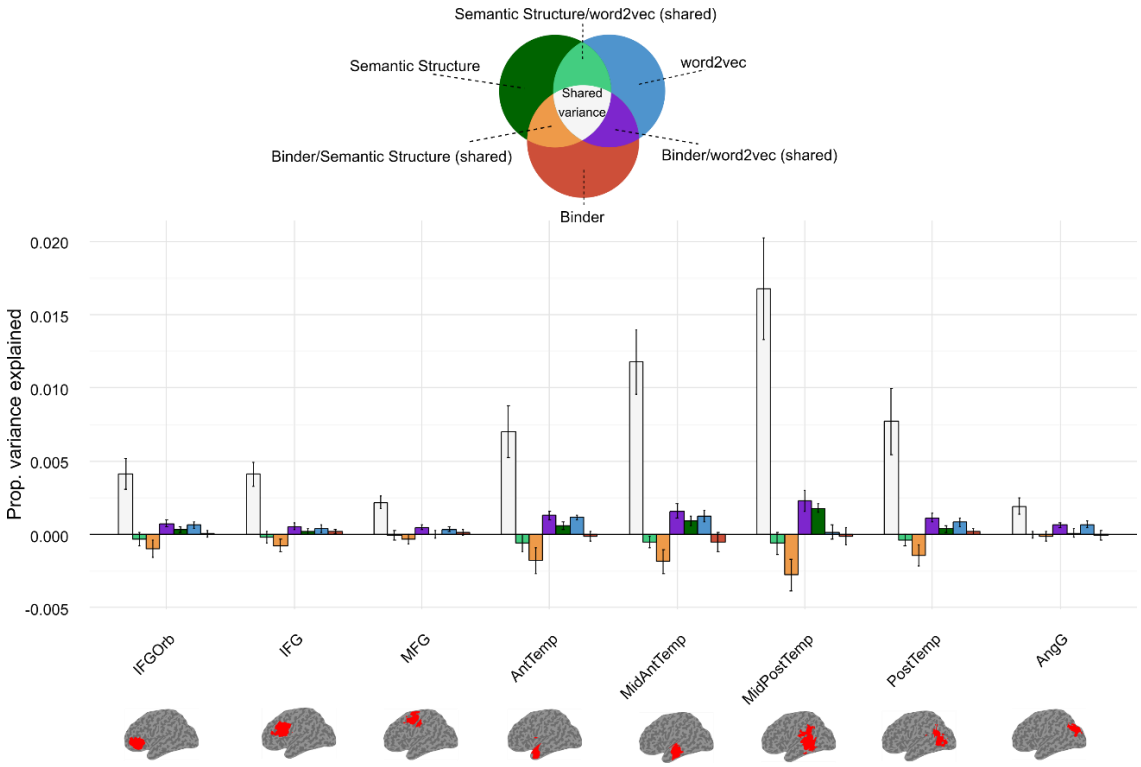


Figure 4.7

Variance partitioning analyses between the Semantic Structure, word2vec, and Binder models. Bars show the mean proportion of variance explained across subjects for each variance partition, in each ROI. Joint

models between pairs of models and the combination of all three models were run, and variance was partitioned using set theory to each single model, pair of models, or the combination of all models, accordingly. Brain maps below ROI labels indicate the parcels used to make subject-specific language-selective ROIs, based on selectivity from an independent functional localizer. Error bars are the standard error of the mean across subjects. The partition with the greatest variance explained was attributable to the shared variance of all three models (white bar). Some individual models or model pairs also showed significant explained variance that could not be accounted for by the other models, including shared Binder and word2vec variance (purple bar), as well as additional unique variance attributable solely to the word2vec model (blue bar) and the Semantic Structure model (dark green). We cannot rule out variance uniquely attributable to the Binder model, as these estimates are not adjusted for downward bias of performing variance partitioning analysis on the held-out validation set. Thus, they are a lower bound on the possible explained variance attributable to each partition, sometimes resulting in impossible values (e.g., the shared variance of Binder and Semantic Structure, orange bar, being negative). Abbreviations: inferior frontal gyrus orbital (IFGOrb), inferior frontal gyrus (IFG), middle frontal gyrus (MFG), anterior temporal (AntTemp), middle anterior temporal (MidAntTemp), middle posterior temporal (MidPostTemp), posterior temporal (PostTemp), and angular gyrus (AngG).

[Manuscript continues with figure on next page]

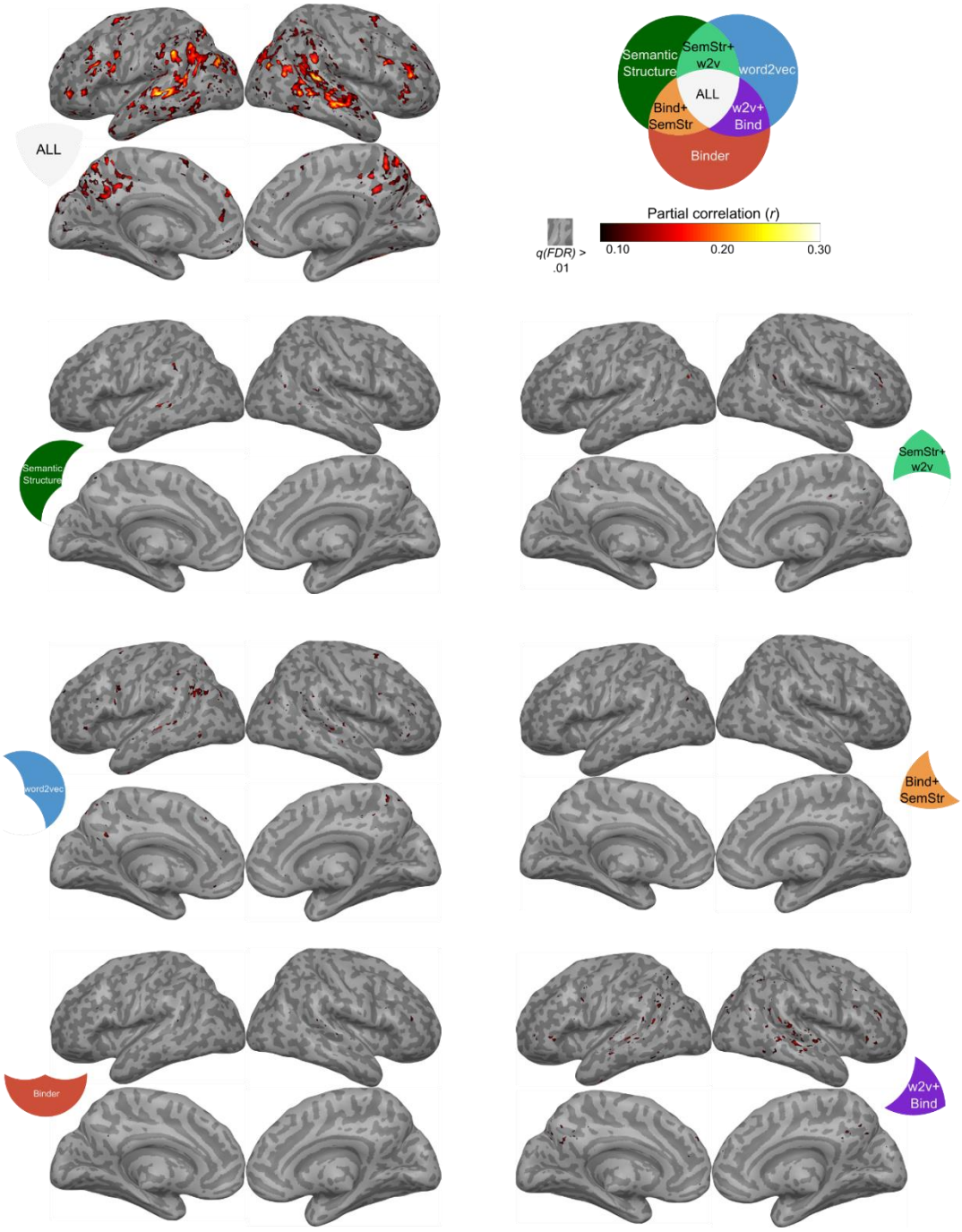


Figure 4.8

Whole-brain maps for variance partitioning analyses between the Semantic Structure, word2vec, and Binder models, for one example subject. Maps are plotted on the inflated cortical MNI surface. The partial correlation is the signed square root of the partial variance explained (r^2) for each of the seven partitions. These maps confirm the ROI analyses presented in Figure 4.7: The majority of variance explained was shared between all three semantic models. Some individual models or model pairs also showed significant explained variance that could not be accounted for by the other models, including Semantic Structure (dark

green) word2vec (blue), and the joint word2vec and Binder models (purple). Note that we cannot rule out variance uniquely attributable to the Binder model alone, as these partition estimates are not adjusted for downward bias of performing variance partitioning analysis on the held-out validation set.

3. General Discussion

The main goal of the current study was to test the degree to which a model of semantic structure based in lexical semantic theory can serve as a candidate model for how the brain represents the meaning of verbs. We also aimed to compare this model to other current leading models of semantics. We accomplished this by using a voxel-wise encoding model approach, in which we predicted fMRI responses to verbs in naturalistic language (audiobooks) in the brain. First, we found that a model based on semantic structure (with features for e.g. Cause, Motion, Contact) showed significant prediction accuracy in language-selective ROIs. We also compared this model with other lexical semantic models and found that the semantic structure model shared substantial explained variance with a model based on sets of syntactic frames, over and above a model based on only the single frame of the verb in context. This confirmed a core prediction of the relationship between linguistic and semantic structure: that the sets of frames a verb does or does not take predicts its event structure. We next compared a semantic structure model with two other leading models: a distributional semantics model (word2vec), and a sensorimotor/cognitive semantic model based on word ratings (Binder). We found that all models resulted in significant prediction accuracy across language-selective cortex, and further, that all three models shared most of their explained fMRI response variance. Nevertheless, there was significant variance attributable uniquely to single or joint sets of models, suggesting these models may capture at least partially non-overlapping semantic information. Together, our findings suggest that a model of semantics based on semantic structure is a plausible model for how the human brain represents the semantics of events, on par with other leading models of semantics.

Our results quantify for the first time the degree to which the predictions of linguistic/semantic structure correspondence from lexical semantic theory are borne out in how the brain responds to verbs in natural language. Additionally, our work is the first to objectively compare the degree to which the semantic structure model compares to other leading models of semantics: distributional semantics (word2vec) and a leading

sensorimotor/cognitive semantic model (Binder). A central finding from this analysis is that information about semantic structure is implicitly present in other previously established model types in semantics (based on the representational similarity analysis of feature spaces) and substantial variance in fMRI responses is shared between the spaces.

3.1. Differences in representations across regions

We found that across all analyses, posterior temporal language-selective regions showed the greatest prediction accuracy, overall. This is consistent with prior literature that suggests that posterior middle temporal cortex may play an especially important role in representing abstract information about events. This region has been shown to respond to event words (both as verbs and nouns) more than other words (Bedny et al., 2008, 2014; Hernández, Fairhall, Lenci, Baroni, & Caramazza, 2014; Peelen et al., 2012; Romagno, Rota, Ricciardi, & Pietrini, 2012). It also appears to represent action categories invariant to incidental visual properties (Hafri et al., 2017; M. Wurm, Caramazza, & Lingnau, 2017; M. F. Wurm & Lingnau, 2015), and across visual and linguistic stimuli (M. F. Wurm & Caramazza, 2018). Although we did not explicitly test for cross-modal information in the current study, an intriguing possibility is that posterior middle temporal cortex represents event semantics across modality. Beyond language-selective areas, our results also revealed other regions with significant prediction accuracy, such as medial parietal areas that have been implicated in memory representations (Binder et al., 2009). Future work should investigate the representational content of regions outside of language-selective areas that represent semantic information.

Somewhat surprisingly, none of the language-selective regions that we investigated showed a benefit for specific syntactic context in which a verb was embedded (the Single Frame model) over the sets of frames the verb takes (the Average Frame model). One of several possibilities may explain such a result. It is possible that models based on explicit syntactic transformations and structure-building must be tested (J. Brennan et al., 2012; J. R. Brennan et al., 2016), rather than the simplified surface frames we implemented here. It is also possible that neural implementation of syntactic context distinctions is not observable in voxelwise fMRI patterns (Blank et al., 2016). Perhaps such differences

could be observable using finer-grained methods (e.g. fMRI adaptation, or recordings at the single unit level). Further experimental work and methodological advances will need to be developed to make progress on this issue.

3.2. Limitations of the current work

Despite the success of the semantic structure approach in the current study, there is room for improvement in its implementation. The database we used to formalize semantic structure, VerbNet, is the most comprehensive dataset available that attempts to operationalize the Semantic Structure Consistency hypothesis (Kipper et al., 2008; Levin, 1993). Its main limitation has to do with the difficulty of identifying the elements of semantic structure: While identifying the sets of frames a verb takes and shares with other verbs is somewhat trivial (as in the Average Frame model), placing a meaningful label on what generalizations the shared frames cognitively correspond to can be challenging, even for seasoned semantic veterans (Fisher et al., 1991). Indeed, we found that a model built from average frames performed better than one with explicit labels of semantic structure features (Figure 4.4), suggesting that the semantic structure labels were imprecise, possibly collapsing over relevant distinctions (e.g., manner and path, Talmy, 2000), or overly differentiating others, such as event roles (Dowty, 1991; White et al., 2017). Nevertheless, some elements of semantic structure for which there is some agreement are Cause (in some form), Motion, State Change, Contact, and Physical/Mental Transfer (Jackendoff, 1990; Pinker, 1989).[14] Furthermore, Hartshorne and colleagues (2014) are beginning to make advances in obtaining psychological judgments of naïve subjects on the classes, frames, and features of large sets of verbs; thus, we can expect improvement here in the near future.

We also ignored effects of sentence processing and narrative construction in modeling the measured fMRI responses. A large literature on sentence processing using behavioral techniques has shown that semantic and syntactic representations are

[14] A question which we do not resolve here is the debate over why such a correspondence may existence between linguistic and these particular elements of meaning (semantic structure) in the first place. Perhaps the linguistic/semantic relationship is transparent precisely *because* such general semantic notions are available conceptually early in development (Strickland, 2016) yet are often opaque to observation alone (e.g. the meaning of a word that refers to an internal mental state, such as *to think*, may be impossible without observation of the sets of syntactic frames such a word appears in (Landau & Gleitman, 1985)).

continuously updated and revised as people comprehend language (e.g., J. C. Trueswell & Kim, 1998). The effects of this on-line linguistic interpretation on when event representations are active, and how these are maintained over time, are unknown. Additionally, how semantic information is maintained and updated over longer periods of time and over longer narratives (e.g. when characters or settings change) is an active field of research not addressed here (Baldassano et al., 2017; J. Chen et al., 2016); our focus was on the individual event representations activated by verbs in natural language. To fully account for semantic representations in the brain, theories and models of sentence processing and narrative structure will ultimately need to be incorporated.

Part of the limitations of this study are of course the measurement limitations which can be considered noise intrinsic to the methodology. A number of known factors contribute to this noise, including the spatial and temporal resolution of fMRI, the intrinsic temporal smoothness of the hemodynamic response, but many unknowns exist as well, such as the exact mapping between neural activity and the fMRI response. Improvements in technology and methods of measurement can reduce the contribution of these factors.

3.3. Implications for semantic compositionality

In the current study, we confined our investigation to the realm of verbs for the purpose of comparing models of their meaning. Thus, we did not test models of semantic composition per se, but rather we use the insight that such composition involves structured representations. However, future work should explicitly address the compositional nature of event representations in the brain. The advantage of semantic structure for investigating compositionality is that it explicitly commits to the existence of different categories with different combinatorial properties: events (one category) are about specific kinds of relations between entities (another category; Williams, 2015). This view is in line with researchers in computational linguistics who have argued that the algorithmic operations of semantic composition should be categorically different depending on the grammatical category of a particular word, and thus that such categories should be explicitly specified before compositional operations take place (Baroni & Zamparelli, 2010; see Kartsaklis, 2014 for review; Mitchell & Lapata, 2010). In contrast, many models (including distributional models and the Binder model) treat all

ontological types the same: objects, events, and attributes are embedded in the same representational space. Thus, the advantage of the semantic structure approach is that it points out interpretable elements of meaning (e.g. Cause), while at the same time differentiating by category. Nevertheless, semantic structure does not have much to say about the semantic representation of the entities involved in events, apart from general attributes (like whether it is a mass/count concept, e.g. *water* vs. *bottles*; singular vs. plural, e.g. *bottle* vs. *bottles*; or animacy, e.g. *animal* vs. *vehicle*; Talmy, 2000a; Jackendoff, 2002). Thus, in future work, models of the semantic representation of entities, whatever they might be, can be integrated with a semantic structure model such as ours for comparison against other semantic models that do not explicitly differentiate entities from events.

3.4. Similarities of semantic models

One of our main findings was that the semantic structure model shared the majority of its explained variance with the other two models of semantics. The first was word2vec (our distributional semantics model). We suspect that this model ends up sharing this variance because its large-scale semi-supervised training is sufficient to approximate structured regularities in syntax made explicit by lexical semantic theories. For example, verbs like *walk* and *run* have a set of canonically permitted syntactic frames, some of which include appearances with prepositional phrases (e.g. *I walk to the park*) and are almost never followed by words like *that* (e.g. **I walk that the park* [ungrammatical]). This is in line with recent work that has found that syntactic frame information can be classified using word vector representations (Kann, Warstadt, Williams, & Bowman, 2019). The other semantic model of interest was Binder and colleagues' sensorimotor/cognitive model. We suspect that this model shares variance with the semantic structure model because some of the features are related (albeit indirectly) to aspects of semantic structure: Cause (semantic structure) is related to Binder's Cause feature; Motion (semantic structure) is related to Binder's Motion feature (*showing a lot of visually observable movement*); and so on. Indeed, Binder and colleagues added such features to their previous iterations of a sensorimotor model (L. Fernandino, Humphries, Conant, Seidenberg, & Binder, 2016; Leonardo Fernandino et al., 2015), in part to capture aspects of abstract words that the sensorimotor model alone did not

capture. With these added features, the Binder model moves closer to an account of semantics that incorporates semantic-structurally relevant features, albeit indirectly.

3.5. The virtue of interpretable models

When several models end up explaining the same variance in fMRI responses to a large degree, as we observed, how is one to choose the "best" model? If the goal is to provide an explanatory model of human cognition, we see a benefit in pursuing models with greater interpretability and plausibility, all else being equal (although of course the line for when to consider models as approximately equal in performance can be debated). The utility of interpretability is that it allows researchers to make contact between linguistic semantics and other conceptual and perceptual systems involved in building an explanatory model of the world, such as high-level vision, memory, or reasoning. In this, the semantic structure approach excels: its elements such as Cause, Motion, State Change are core concepts that are available early in development, before 9 months of age (Kominsky et al., 2017; Leslie & Keeble, 1987; Muentener & Carey, 2010), and are hypothesized to act as the foundation for further conceptual development (Carey, 2009; Strickland, 2016). Further, some of these high-level representations are available as part of perceptual processing itself (Hafri, Trueswell, & Strickland, 2018; e.g. causality and agency; Rolfs et al., 2013). Another criterion is plausibility: the models we investigated make different assumptions about how the relevant semantic information is learned and organized. Distributional semantic models achieve their unparalleled performance via training on large text-based datasets, often with hundreds of millions of examples or more. However, by age 4 children know thousands of verbs and their meanings, yet they do not get exposed to near the amount of linguistic input that appears to be necessary for word2vec models to succeed: Hart & Risley (1995) estimate that children hear only 13-45 million words by age 4 years. Indeed, the evidence in the psycholinguistics community suggests that humans do not engage in a pure associative learning procedure for learning the meanings of words. Instead evidence points to a highly inferential process based on few instances (J. C. J. C. Trueswell, Medina, Hafri, & Gleitman, 2013; Woodard, Gleitman, & Trueswell, 2016). This includes the semantic structure of verbs, as observed in studies of verb-learning based on linguistic and perceptual input (Gleitman, 1990; Landau & Gleitman, 1985; Yuan & Fisher, 2009).

Thus, even if distributional semantic models end up at an approximately similar semantic state as humans, the learning processes are likely qualitatively different.

This is not to discard entirely the utility of distributional semantic models such as word2vec for offering insights into the learning and organization of human semantic knowledge. Indeed, the fact that they perform so well at approximating human performance in semantic tasks (Mandera et al., 2017) and predict a large amount of brain response variance to language (Huth et al., 2016; Pereira et al., 2018) suggests that their representational space may offer insights into the kinds of information latent in word distributions in language. Additionally, comparison of model architectures (such as semi-supervised models like word2vec vs. LSTM models) and their relationship to human language processing or representation can also offer insight (Linzen, Dupoux, & Goldberg, 2016). Such work has also gained traction in the vision community using deep convolutional neural networks (Bonner & Epstein, 2018). Other approaches involve a "hybrid" approach, attempting to take insights from both distributional semantics and compositionality (Fyshe et al., 2014; Fyshe, Wehbe, Talukdar, Murphy, & Mitchell, 2015; Lenci, 2018). We note, however, that investigating how such models might inform our understanding of human cognition requires understanding the computational problem at hand (semantic understanding), the kinds of representations that are likely to exist (e.g. in this case, semantic structural representation), and the feasibility of the models as candidates for human learning and representation (Marr, 1982). To gain any high-level interpretation from it requires knowing what one is looking for (Krakauer, Ghazanfar, Gomez-Marin, Maciver, & Poeppel, 2017). In contrast, both Binder and the Semantic Structure model make explicit different aspects of meaning: event structure in the former, and sensorimotor and cognitive associations in the latter.

3.6. The importance of model comparison

Our results reveal the importance of explicit model testing and comparison. Notably, previous work using distributional semantic models (Huth et al., 2016; Pereira et al., 2018) have not reported such explicit model comparison. Without this, it is difficult to make claims about the superiority of one model over another, and to identify the precise sharing of information between them (both in information shared among features, and in explained variance in the brain). In future work, we can also use explicit model

comparison to infer exactly *how* the models capture similar variance, and what features map across models, e.g. by performing canonical correlation analysis (finding the mapping between two feature sets maximizing their correlations). By identifying the commonalities across these models, and improving each, we will make progress toward a full explanation of semantic representation in the human brain (e.g., Lescroart & Gallant, 2018). Of course, this does not discount the usefulness of carefully controlled experiments. We believe our approach is important and complementary to the precise experimental control afforded by single sentence stimulus manipulations (e.g. for investigating syntactic alternations). However, to distinguish between high-performing models of how the brain responds to language may ultimately require this careful control to test a critical condition by which to distinguish candidate models.

3.7. Conclusions

The central contribution of our work is that properties of event structure (e.g. Cause, Contact) are encoded spontaneously by people listening to naturalistic speech. Further, it is proof of concept that we can make contact between theories of lexical semantics and data-driven statistical approaches to language. Our findings support the theory that an important – but not the only – aspect of how the brain represents semantics is via semantic structure: that is, the nature of relations between entities. More broadly, our findings suggest that progress can be made in modeling brain responses to language using interpretable, theoretically meaningful semantic properties. Although just in its beginnings, the semantic structure approach is a step towards a core component of the human semantic capacity: to build meaning through composition.

4. Supplementary Methods

4.1. Participants

Six adults were recruited from the University of Pennsylvania community (4 female, 2 male; ages 20, 25, 26, 28, 31, 33). All participants were healthy, had normal or corrected-to-normal vision and normal hearing, and provided written informed consent in compliance with procedures approved by the University of Pennsylvania Institutional Review Board. Four were right-handed and two were ambidextrous. All were native English speakers, and one was also fluent in Arabic. Data from an additional participant

(male, 22 y.o.) was discarded before analysis for an inability to complete the entire experiment due to constant drowsiness. Scanning took place over the span of three to four days, with scan sessions lasting 1.5-2 hours for each visit. Note that our goal was to collect a large and reliable set of data for each subject, rather than a small set of data for many subjects, as is standard for the voxelwise encoding model approach (Huth et al., 2016; Lescroart & Gallant, 2018; Nishimoto et al., 2011). The pattern of results was nevertheless consistent across subjects. Subjects underwent additional scans of six 10-minute silent film clips and additional functional localizer scans that will not be reported in this manuscript.

4.2. fMRI data collection

Scanning was performed at the Center for Functional Imaging at the University of Pennsylvania on a 3T Siemens Prisma scanner equipped with a 64-channel head coil. High-resolution T1-weighted images for anatomical localization were acquired using a 3D magnetization-prepared rapid acquisition gradient echo pulse sequence (MPRAGE, repetition time [TR], 2200 ms; echo time [TE], 4.67 ms; flip angle, 8°; voxel size, 0.94 × 0.94 × 1 mm; matrix size, 192 × 256 × 160 mm]. T2*-weighted images sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired using a multiband gradient echo echoplanar pulse sequence (TR, 2000 ms; TE, 25 ms; flip angle, 70°; voxel size, 2 × 2 × 2 mm; multiband factor, 3; matrix size, 96 × 96 × 81). Field mapping was performed at the end of each scan session with a dual-echo (echo time (TE) = 4.06, 6.52 ms) gradient echo sequence with pulse repetition time (TR) = 1200 ms, flip angle (FA) = 60°, pixel bandwidth = 260, voxel size, 3.4 × 3.4 × 4.0 mm; matrix size, 220 × 220 × 208 mm. Phase difference and magnitude data were saved from each channel.

Visual stimuli for non-audiobook runs were displayed at the rear bore face on an InVivo SensaVue Flat Panel Screen at 1920 × 1080 pixel resolution (diagonal = 80.0 cm, width × height = 69.7 × 39.2 cm). Participants viewed visual stimuli through a mirror attached to the head coil. Responses for functional localizer scans were collected using a fiber-optic button box. Sounds were presented through MRI-compatible earbuds (Sensimetrics S14) in-ear piezo-electric headphones (Sensimetrics S14). These headphones provide high-quality audio and attenuation of scanner noise. Audio stimuli were pre-processed to account for the resonance properties of the earbuds and were

presented at comfortable levels.

4.3. Stimuli & Task

Fiction book excerpts (both text and audio) were selected from freely available online databases: gutenberg.org (free eBooks) for text, and archive.org for audio. Audiobooks were generally recorded by amateur speakers. We extracted book content from the start of each book until a natural stopping point (most often chapter end) such that we had between 5-15 minutes of content for each. Book choice was based on the following criteria. First, only books with high audio quality and speaker engagement were selected. We also used several objective criteria: minimize amount of dialogue (number of verbs like *say, tell*); maximize concreteness of verbs and other words (based on ratings in Brysbaert, Warriner, & Kuperman, 2014); minimize mean length of utterance, a proxy for sentence complexity. We chose books that we expected would be unfamiliar to participants (this was confirmed; on average, only 3 books were very familiar to each subject). Final choices were made from books that met the above criteria by maximizing the variety in book genre/content (e.g. science-fiction, fantasy, fairy tale) and speaker accent/gender. The final set featured 8 American English speakers (4 male), 1 Australian (male), and 5 British (3 male). Final book choices were: Armageddon—2419 A.D. (Philip Francis Nolan); Into the Wild (Jon Krakauer); Black Beauty (Anna Sewell); The Awakening (Kate Chopin); The Jungle Book (Rudyard Kipling); The Tale of Peter Rabbit (Beatrix Potter); The Lion, The Witch, and The Wardrobe (C.S. Lewis); The Monster (Stephen Crane); The Cosmic Computer (H. Beam Piper); The Invisible Man (H.G. Wells); The Wind in the Willows (Kenneth Grahame); The Fish and the Ring (Joseph Jacobs); Jack and Jill (Louisa May Alcott); The Golden Bird (Brothers Grimm). For these final books, audio quality was further improved using filtering and equalization tools (e.g. noise cancellation) in Audacity software (www.audacityteam.org). All audio was sampled at 44.1 kHz. Five of these books were selected as the validation stimulus set and were repeated once (the first five in the list above); the other nine were used as the estimation stimulus set and were only played once. Ten seconds (5 TRs) of silence were added to the end of each stimulus account for the delay in hemodynamic response to the end of the auditory stimulus. In the scanner, participants were instructed to attend to the narrative of the audiobook. A light-gray screen with centered crosshair was visible

throughout all scans.

4.4. Construction of feature spaces

To construct feature spaces, it was first necessary to identify the precising timing of each word onset/offset. To do this, we aligned the audio with the text for each book. First, text was corrected to conform to the exact words used in the audio. Then the text and audio were aligned at millisecond precision using the Penn Phonetics Lab Forced Aligner software (http://fave.ling.upenn.edu). The software is an automatic phonetic alignment tool that uses hidden Markov models to identify the start and end of phonemes and words. After alignment, the timings were verified in the phonetics software Praat (www.praat.org).

For each feature space, we modeled features of each instance of each verb in the stimulus set. To do this, we first used automatic part-of-speech (POS) tagging as implemented by the Stanford POS tagger (https://nlp.stanford.edu/software/tagger.shtml). The bidirectional version of the tagger is about 97% accurate, and subsequent to tagging, we verified and corrected POS tags to ensure we identified all verbs. We excluded the following categories of verb types: auxiliaries (e.g., *will/shall, have,* and *do* as in *I will go, I have seen it*, or *Did you know?*); modals (*can, may, must, should,* etc.); and the copula (*be*, as in *I am fine*). Present and past participles were included in the analysis (e.g. *The man running on the track was happy; The snow, melted by the sun, was cold.*). Only verbs that appeared in both word2vec and VerbNet (see below) were included (for Binder, we collected our own ratings on the verb set). After pre-processing, verbs made up approximately 15% of all words in the stimulus set. Verbs were then lowercased and lemmatized (i.e. converted to word roots, e.g. *gave → give, giving → give*).

4.4.1. Word2vec features

Word2vec is an implementation of the distributional semantic approach that has achieved remarkable success at semantic tasks (Mikolov et al., 2013) and at accounting for semantic priming (Mandera et al., 2017); similar models have also predicted cortical responses to language (Huth et al., 2016). In this model, each word is represented as a vector embedded in a multidimensional space (in this case, 300 dimensions). Vectors for each word are learned via a shallow neural network that are trained to predict a word's

context (neighboring words) given an input word (the skip-gram architecture) or vice versa (continuous bag-of-words, or CBOW). After training, the embedding for each word is its corresponding weights to the hidden layer of the neural network (ranging from -1 to 1). Words that share similar contexts share similar word vectors. Note that precise order of words in the context is not given special status, although in skip-gram, closer words are weighted more heavily. We used Google's pre-trained word2vec model, which was trained on a corpus from Google News of about 100 billion words, with a context window of size 10 and dimensionality of 300. See Mikolov et al. (2013) for more details on the model training and performance. The choice of this model was based on its high performance and the convenience of its availability, but other distributional semantic models would be expected to perform similarly. Indeed, work has shown that different distributional semantic approaches are highly dependent on the hyperparameters chosen and on amount/content of training data; despite differences in their training algorithms, other distributional semantic model types (e.g. Latent Semantic Analysis) can perform similarly given the right training (Levy, Goldberg, & Dagan, 2015).

To extract features for each verb from word2vec, we simply extract the 300 feature values for each verb in the stimulus set. Note that these values will always be the same, no matter the context of the verb or its sense.

4.4.2. Verb Class, Syntactic Frame, and Semantic Structure features

To extract features of verbs relevant for lexical semantic theory, we used a database called VerbNet (https://verbs.colorado.edu/verb-index). VerbNet is a publicly available database in which verbs are organized into *Verb Classes* based on the sets of syntactic frames a verb canonically does or does not take (see Introduction for further explication of the basis for such an organization). The database is an extension of Levin's (1993) work on verb classes by Kipper-Schuler and Palmer (Kipper et al., 2008) to include more verbs and refine the classes. VerbNet contains over 9,000 verb entries, and over 300 Classes. Of the 744 unique verbs that appeared in our stories, 675 (90.7%) appear in VerbNet, resulting in 93% of verb tokens (instances) with existing values in VerbNet.

In VerbNet, each verb class has a list of syntactic frames associated with it. Each frame is a possible syntactic frame that verbs in the class can take. The frame is a surface representation that includes the number of phrase types before and after the verb (e.g., Noun Phrase, Verb, Noun Phrase, Prepositional Phrase, i.e. NP V NP PP); common

syntactic transformations such as the passive (movement of object to subject position) are not separately described. Note that specific syntactic frames repeat across verb classes. For example, the NP V NP (transitive) frame is very common and appears across many verb classes. For each syntactic frame, a semantic structure description is given that includes semantic predicates: Boolean semantic elements that take aspects of the event as arguments, including thematic roles (e.g., Agent, Patient, Theme). As discussed above in Methods, the same syntactic frame (e.g. NP V NP) will be associated with different semantic structure elements; for example, for a verb like *break*, a semantic feature like Cause will be associated with NP V NP, while for a verb like *see*, a feature like Perceive will be associated with NP V NP. See Figure 4.3 for more concrete examples. A coarse temporal structure of the event is also specified (e.g., whether Motion or Contact occurred at the beginning of the event); however, we ignore this aspect of the semantic representation here. VerbNet is being continuously updated for more verb coverage. See above website and references for more information on VerbNet's implementation of semantics and classes, as well as its current state.

An example of a syntax/semantics entry for the verb *hit* is the following (E stands for an "event variable" that allows temporal order to be specified):

- Verb Class: *hit-18.1*
- Syntactic Frame: NP V NP (Noun Phrase, Verb, Noun Phrase)
- Example Sentence: Jimmy hit the table.
- Semantic Structure: Cause(Agent, E), Manner(during(E), directed-motion, Agent), NOT Contact(during(E), Agent, Patient), Manner(end(E), forceful, Agent), Contact(end(E), Agent, Patient)

Figure 4.3 in main text gives a visual illustration of how features for each model correspond to example sentences in each context. For the Verb Class model, the features were simply the verb class as annotated in context (e.g. for the verb *drop* in *Jimmy dropped his ice cream*, *roll-51.3.1*). For the Single Frame model, the features were the phrase type pre- or post-verb (e.g., first noun phrase [NP] pre-verb, second NP pre-verb, first NP post-verb, first prepositional phrase post-verb, etc.). For the Average Frame model, the features were the same as for the Single Frame model, except the features were average over all frames within the verb's annotated class. For example, if a post-verb NP appears for the verb *drop* in its labeled class (e.g. *roll-51.3.1*) in 4 out of 8

frames, this feature would receive the value 0.5, regardless of the particular syntactic context of the instance of the verb. For the Semantic Structure model, the features were the semantic predicates for the particular frame (e.g., Cause, Motion, Contact, NOT Contact), as well as the event roles that were arguments of those predicates (e.g., Agent, Patient, Theme, Topic). If a feature appeared twice in this layer (e.g., Agent in Cause and Agent in Contact, as in the example above), it was only coded once. The temporal predicates (e.g., start, during, end) and predicate type features (e.g., forceful, directed-motion), as in the above example, were not included in the model.

We primarily used version 3.2 of VerbNet, as this was the version available at the time the project was initiated. Version 3.3 of VerbNet became available mid-way through the project, and although it had greater coverage, the system of semantic structure features was changed. However, to augment the verb coverage for this study, we used VerbNet 3.3 coding for any verb that did not appear in VerbNet 3.2.

Each verb instance in the stimulus set needed to be labeled with its particular verb class and frame in context. However, there are not yet reliable methods to do so for this database. Thus, the first author annotated each verb instance by hand, using custom annotation software written in python with the nltk package. The software presented each verb one at a time in its sentential context, along with the possible frames for each verb class that the verb matched. Each frame also presented the associated example sentence in the VerbNet database. (The semantic predicates were not presented, so as not to bias choice based on these properties.) VerbNet contains entries for some verb-particle constructions (e.g., *look after* meaning to care for), so these entries were displayed as possible choices as well. If an entry was a close but not exact match semantically and syntactically, the closest matching class/frame was chosen. If the usage of the verb in the particular context did not match syntactically or semantically, the verb was skipped and received no annotation (and was therefore excluded from modeling). This annotation procedure was performed for all verbs in the stimulus set, 3,146 in total.

For modeling purposes, we only wanted to include features that appeared at some minimum frequency that we could reliably estimate its contribution to fMRI responses. To this end, we only included a feature if it appeared at least 3 times in the estimation stimulus set, and 2 times in the validation stimulus set (since these runs were repeated once). After this feature culling, the following is the number of features remaining for

modeling, for each model: Verb Class, 98 out of 356; Single and Average Frame, 18 out of 37; Semantic Structure, 99 out of 232. Features and frequency for each model are listed below, in Figure 4.9.

[Manuscript continues with figure on next page]

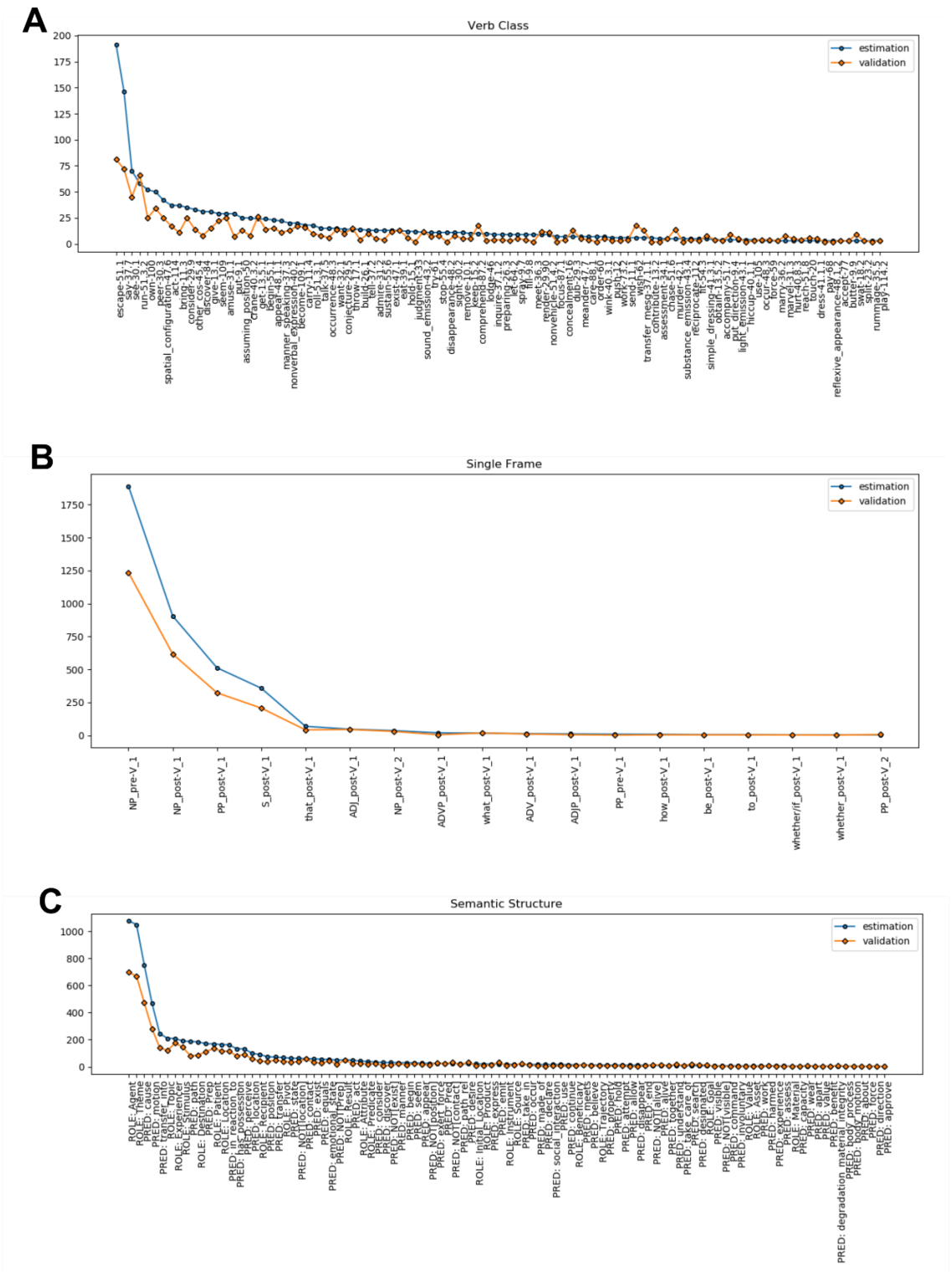


Figure 4.9
Frequency of features for each model based on lexical semantic theory. Frequencies are listed separately for

both estimation and validation stimulus sets. Only features frequent enough to be included in modeling were used (minimum 3 in training set, 2 in validation set). Each feature set is ordered by estimation set frequency. **A.** Verb Class features (98 total). Details about each Verb Class (i.e. its members, frames, and semantic structure) can be found in the online VerbNet database. **B.** Single Frame features (18 total). The same set of features were also used for the Average Frame model, except features in that model were averaged across all frames within Verb Class. Uppercase letters represent phrase types or part of speech of single words (i.e., NP, Noun Phrase; PP, Prepositional Phrase; ADJ, adjective; ADV, adverb; S, sentence complement). Lowercase letters represent specific words (e.g., *that, how*). Pre-V indicates the word/type appears before the verb in the surface syntactic frame, post-V after. Number indicates whether it is the first or second of that type appearing before or after the verb (e.g., *NP_post-V_2* indicates that the feature represents the second of two NPs post-verb in a frame). **C.** Semantic Structure features (99 total). The text "ROLE:" precedes semantic role features (e.g., Agent, Patient), and "PRED:" precedes semantic predicate features (e.g., Cause, Contact).

4.4.3. Binder features

As an alternative model to our Semantic Structure model based on VerbNet features, we also tested the sensorimotor/cognitive model of Binder et al. (Binder et al., 2016). The original version of this model includes 65 semantic features corresponding to different cognitive domains: for example, sensory (visual, auditory, etc.), motor, spatial, event, cognition, and emotion. It has been used successfully to account for fMRI responses to single sentence stimuli across the brain (Anderson et al., 2017, 2018). Binder et al.'s dataset includes average ratings for 535 English words: 434 nouns, 62 verbs, and 39 adjectives.

To use their model requires that we have data for all the verbs that appear in our dataset, only 47 of which appeared in Binder et al.'s set. Thus, we collected our own ratings on Amazon Mechanical Turk for our 675 verbs. We chose a subset of the full feature set for data collection, as Binder et al. found that not all features were needed to explain variance across their dataset: in a factor analysis (across the nouns and verbs), they found that 16 factors accounted for 81% of the variation across words/features.

The subset of features used were chosen through several criteria, based on the results of factor analysis (described previously in Binder et al.). First, in order that the features chosen would reflect the underlying latent factor to a reasonable extent, only features with a unique loading of at least 0.50 on a factor were considered. Next, from this set, one feature was retained for each factor. This was either the feature with the top loading value, or a similarly high loading substitute that we believed would capture a more general theoretical property of the factor. For example, Factor 3 grouped together the features (in order) Sound, Audition, High sound, Loud, Low sound, and Music. Instead

of choosing Sound ("being associated with a characteristic or recognizable sound"), we chose the more general Audition ("being associated with hearing something"), and in all cases, the features that were substituted had similarly high loadings. Finally, we added additional factors for theoretical considerations:

- To include features that queried use of additional sensory modalities beyond Audition and Smell, we included Vision and Touch;
- To include an additional feature from the motor domain, we added LowerLimb;
- To include visual features that likely corresponded to features in our Semantic Structure model, we added Motion;
- To include cognitive features that likely corresponded to features in our Semantic Structure model, we added Cognition, Caused, and Social

After this selection, we were left with 23 of the 65 features for rating, listed below in Table 4.1.

Eight ratings were obtained for each verb for each query. MTurk participants were presented with past tense forms of six verbs from the list (no sentential context), one at a time, and rated each on a scale of 0 to 6 on each query. Data from participants whose correlation with the mean for other participants in their set of verbs was less than 0.5 were discarded. To confirm that our rating procedure produced similar values as that of Binder et al., we checked the correlation between the shared set of words in their study and our study (47 verbs total). The average correlation for each word across the queries used in common was 0.88, indicating that our rating procedure was sufficient to produce similar results, despite the lower sample size (Binder and colleagues collected 30 ratings per word).

Table 4.1

Name	Type	Modality	Submodality	Query (To what degree do you think of this as…)
Vision	Sensory	Vision	General	being an action or activity in which you see something
Pattern	Sensory	Vision	Surface	being associated with a visual surface pattern or change in a visual surface
Motion	Sensory	Vision	Motion	being associated with a specific type or a large amount of visible movement
Slow	Sensory	Vision	Motion	being associated with slow visible movement

Face	Sensory	Vision	Shape	being associated with visible movements of the face
Touch	Sensory	Somatic	General	being an action or activity in which you feel something by touch
Audition	Sensory	Audition	General	being associated with hearing something
Smell	Sensory	Olfaction	Quality	being associated with smelling something
UpperLimb	Motor	Motor	Motor	being an action or activity in which you use the arm, hand, or fingers
LowerLimb	Motor	Motor	Motor	being an action or activity in which you use the leg or foot
Landmark	Spatial	Navigation	Navigation	being an action or activity in which you use a mental map of your environment
Path	Spatial	Navigation	Navigation	being associated with someone or something moving from one location to another
Number	Number	Number	Number	being associated with a specific number or amount
Duration	Event	Temporal	Duration	being an action or activity that has a predictable duration, whether short or long
Caused	Event	Causal	Causal	being associated with someone or something causing a change in something else
Social	Event	Social	Social	being an action or activity that involves an interaction between people
Communication	Cognition	Social	Communication	being an action or activity by which people communicate, or transmit or receive information
Cognition	Cognition	Cognition	Cognition	being a mental activity or state of mind that involves thinking
Benefit	Evaluation	Cognitive	Positive	being an action or activity that could help or benefit you or others
Pleasant	Evaluation	Affective	Positive	being an action or activity that you find pleasant
Unpleasant	Evaluation	Affective	Negative	being an action or activity that you find unpleasant
Surprised	Emotion	Neutral	High	being associated with feeling surprised
Needs	Drive	Basic	Basic	being an action or activity that provides things that would be difficult to live without

4.5. Feature extraction and preprocessing

Word features were sampled at 16 Hz. Feature values were initialized at zero, and verb features were inserted into the feature timecourse at the temporal mid-point of each word. To align each feature space with the timecourse of the fMRI data, each feature was downsampled to the fMRI acquisition rate (0.5 Hz) using Matlab's decimate function

(first low-pass filtered with an 8th order Chebyshev Type I lowpass filter, and then resampled). To account for the delay and temporal smoothness of the hemodynamic response function, each feature space included finite impulse response (FIR) predictors for each downsampled feature at each of four delays: 2, 4, 6, and 8 seconds. To place all features of all models on similar scales for modeling purposes, feature timecourses were *z*-scored within run.

4.6. fMRI data pre-processing

After acquisition, data were organized according to the BIDS neuroimaging specification, designed to promote consistent description and organization for data sharing and collaboration (Gorgolewski et al., 2016). All data (audiobooks and localizers) were then preprocessed using FMRIPREP version 1.0.14 (Esteban et al., 2019), a robust and automated pre-processing pipeline for fMRI data. In this pipeline, each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1, and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL v5.0.9).

Functional data was slice time corrected using 3dTshift from AFNI v16.2.07 and motion corrected using mcflirt (FSL v5.0.9). Distortion correction was performed using an implementation of the TOPUP technique using 3dQwarp (AFNI v16.2.07). This was followed by co-registration to the corresponding T1w using boundary-based registration with 9 degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, field distortion correcting warp, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation. For more details of the pipeline see https://fmriprep.readthedocs.io/en/latest/workflows.html. After data were

pre-processed with FMRIPREP, they were high-pass filtered (100 sec) using FSL to remove low temporal frequencies due to scanner signal drift. They were then smoothed with a 3mm FWHM Gaussian kernel, separately within voxels of the same tissue type, using FSL's SUSAN algorithm.

All analyses in this manuscript were conducted in subject space, within a mask consisting of the union of gray matter masks of all subjects and the MNI template, dilated with a 3mm FWHM Gaussian kernel. Whole-brain maps for Figure 4.6B and Figure 4.8 were transformed from subject to MNI space using nonlinear registration warp matrices generated from FMRIPREP pre-processing, and then the maps were projected onto the MNI cortical surface using the Matlab mni2fs toolbox (https://github.com/dprice80/mni2fs).

4.7. Functional localizers and definition of ROIs

We constrained our main analyses to a set of regions in the brain that are known to selectively respond to language. Since the exact locations of these regions varies somewhat from individual to individual, we took the functional localization approach, first used for studies of high-level vision (Kanwisher, 2010), and subsequently brought into the domain of high-level functions of language (Fedorenko et al., 2010). The contrast used to identify language selectivity is written or spoken intact language (semantically and syntactically coherent sentences) compared to a perceptually matched input with degraded language (in the written domain, strings of nonsense words; in the spoken domain, filtered speech such that interpreting coherent words and syntactic structure is not possible).

In the current study, we used the spoken language contrast (Scott et al., 2016). In two six-minute runs, subjects listened to 16 18-second blocks, half of intact and half of degraded speech (see Scott et al for details). We used general linear models (GLMs) implemented in FSL（http://fsl.fmrib.ox.ac.uk/fsl/fslwiki) to estimate the response of each voxel to each speech condition in each scan run. To define ROIs in each subject, we used a group-constrained subject-specific (GSS) ROI definition method (Fedorenko et al., 2010; Julian et al., 2012). This approach yields similar individual subject functional ROIs to the traditional hand-drawn ROI pipeline but uses an objective and automatic method. Anatomical parcels were defined in a large group of subjects undergoing a

similar contrast using a watershed algorithm to identify areas where subjects shared selectivity. We defined each subject-specific ROI as the top 200 voxels in each hemisphere that responded more to the contrast of interest and fell within the group-parcel mask for the given ROI. Although language function has been observed to be primarily left-lateralized in fMRI activation studies and patient studies (Blank & Fedorenko, 2017; Fedorenko & Varley, 2016), in other studies using the encoding model approach (Huth et al., 2016), similarity across hemispheres was observed. Thus, our ROIs were bilateral.

4.8. Reliability of fMRI responses to audiobooks

To determine response reliability across the brain and in ROIs, we calculated the Pearson correlation of voxel timecourses between the repeats of the audiobooks in the validation set (concatenated; 1686 timepoints total). This analysis will result in high correlations if the voxel responds consistently to any aspect of the stimulus (low or high level). Average reliability and proportion of significant voxels ($q(FDR) < .05$) for each ROI and the rest of cortex are reported in Table 4.2. These results demonstrate that a significant proportion of voxels in our ROIs demonstrated reliable signal. Thus, these data can be used for testing our voxelwise encoding models.

Table 4.2.

ROI	Number of voxels	Mean reliability	Proportion significant
IFGOrb	400	0.068 (0.013)	0.592 (0.114)
IFG	400	0.083 (0.014)	0.670 (0.118)
MFG	400	0.071 (0.010)	0.586 (0.093)
AntTemp	400	0.091 (0.016)	0.716 (0.124)
MidAntTemp	400	0.130 (0.018)	0.830 (0.091)
MidPostTemp	400	0.168 (0.030)	0.876 (0.092)
PostTemp	400	0.109 (0.024)	0.720 (0.146)
AngG	400	0.086 (0.018)	0.660 (0.125)
Non-ROI voxels (all)	149736 (5441)	0.028 (0.004)	0.273 (0.051)
Non-ROI voxels (reliable only)	40718 (7578)	0.081 (0.002)	1.00

Reliability (Pearson r) of fMRI responses to repeats of validation stimulus set. Results listed for each ROI, for all non-ROI voxels, and for non-ROI voxels that were significantly reliable. Significant was set at q(FDR) < .05. Mean ± SE across subjects (no SE for number of voxels in ROIs, as ROIs contained the same number of voxels for all subjects). Abbreviations: inferior frontal gyrus orbital (IFGOrb), inferior frontal gyrus (IFG),

middle frontal gyrus (MFG), anterior temporal (AntTemp), middle anterior temporal (MidAntTemp), middle posterior temporal (MidPostTemp), posterior temporal (PostTemp), and angular gyrus (AngG).

4.9. Voxelwise modeling

The main goal of this study was to test how well each model of interest could predict fMRI responses to language. To this end, we fit voxelwise encoding models to our estimation data set and tested their predictive accuracy on a held-out validation data set. Because of the large number of features relative to observations, we used ridge regression, which trades a small degree of bias in beta estimates for a large reduction in variance of the estimates, as has been used previously in the encoding model approach (Huth et al., 2016; Lescroart & Gallant, 2018; Nishimoto et al., 2011). First, to place all variables and fMRI responses on the same scale, each voxel timecourse and was *z*-scored within scan run (feature timecourses were already *z*-scored within run during feature pre-processing). Additionally, we included several nuisance regressors designed to capture low-level responses to the stories that were not relevant for high-level semantic encoding. These were included at model estimation but were not used to generate response predictions for the validation set. The nuisance regressors included were auditory envelope (A-weighting filter, available at https://www.mathworks.com/matlabcentral/fileexchange/46819-a-weighting-filter-with-matlab; 1 feature); word rate and phoneme rate (i.e., number of words and phonemes occurring within a timepoint); and part of speech for nouns and verbs (i.e., a binary coding at the midpoint of a noun or verb, only for verbs that were modeled; 2 features). These regressors were pre-processed in the same way as the model features of interest.

For each model tested, we first estimated model beta weights on the estimation set (nine audiobooks, 2,632 timepoints total). We performed leave-one-run-out cross-validation within the estimation set to find the best ridge parameter for fitting to the validation set. On eight of nine runs, we tested a range of 20 ridge parameters (zero [standard least-squares regression] and 19 other values log-spaced between 10^0 and 10^4) for each voxel. The lambda that resulted in the highest mean prediction accuracy (Pearson *r*) across cross-validation iterations was chosen, separately for each voxel.

To generate timecourse predictions on the validation set, the full estimation set was used to estimate beta weights for each feature, for each voxel. These weights were used to

generate predicted responses to the validation set data according to the stimulus features of the validation set. The validation stimulus set consisted of five audiobooks repeated once (1,686 timepoints total). fMRI data were averaged across repeats to improve reliability. The measure of prediction accuracy we used was the Pearson correlation between the predicted and actual response, per voxel. For ROI analyses, the mean prediction across voxels within each ROI was calculated (per subject). For whole-brain maps, FDR-correction was applied to correlation values across voxels, with the significance threshold set at $q(FDR) < .05$.

For all models (including joint models for variance partitioning, below), we used a generalized form of ridge regression called Tikhonov regression (Tikhonov & Arsenin, 1977). In our case, this was implemented like standard ridge regression, but instead of the same ridge parameter applied to all features (e.g., in a joint Binder + word2vec model, the same ridge parameter for Binder features and word2vec features), all models were permitted their own ridge parameter (e.g. Binder features would have one ridge parameter, and word2vec another). In recent work using voxelwise encoding models, this approach ("banded" ridge regression) improved prediction accuracy and variance partitioning for joint models (Nunez-Elizalde, Huth, & Gallant, 2018). Nuisance features were treated as one model here (i.e. they collectively received one ridge parameter of their own). For cross-validation, this "banded" ridge approach entailed searching for the best combination of ridge parameters (one for each model). However, for every additional feature space, the search space increases exponentially (2 models: 20^2 possible ridge parameter combinations; 3 models: 20^3; etc.). We found in practice that we could achieve almost identical results to searching the full parameter space by picking the best lambda from each model combination. For example, for a joint model of feature spaces A, B, and C, we could choose the best parameter from pairwise combinations of models (best parameter for A from A+B and A+C combination; best for B from A+B and B+C combination; and best for C from A+C and B+C combination). This drastically decreased computation time from searching across 20^3 parameters to 2^3; and so on for different numbers of models.

4.10. Representational Similarity Analysis across model feature spaces

To test the relationship between model feature spaces, we performed a representational similarity analysis of pairwise feature spaces (Kriegeskorte et al., 2008). For each model of interest (word2vec, Binder, and Semantic Structure), we constructed matrices of timepoints (4,318 in total) by features, after feature pre-processing (downsampling and *z*-scoring within run). We then we calculated the pairwise squared Euclidean distance of each observation to every other observation. This produced a timepoint by timepoint representational dissimilarity matrix (RDM), representing the pairwise dissimilarity of each timepoint to every other timepoint for each model. We compared each model to every other model by calculating the Pearson correlation of the lower off-diagonal values of their RDMs, which produced an estimate of how similarly each model represents our stimuli at the acquisition rate of fMRI. We also performed a similar analysis on matrices of verb instances by features (*z*-scoring across features and then instances), which yielded similar correlation values.

4.11. Variance partitioning analyses

To conduct variance partitioning analyses, we fit models with concatenated feature spaces (see main Methods for details on the models fit). Separate ridge parameters were found for each feature space via leave-one-run-out cross-validation within the estimation data set (see above). After models were fit and predictions made for each joint model, prediction accuracy was converted to variance explained by squaring the Pearson correlation value while retaining its sign. We then used set theory to find the shared and unique variance for each partition, according to the following equations, where each letter A through C corresponds to a different model (equations after de Heer et al., 2017; Lescroart & Gallant, 2018):

For two feature spaces:

1) $r^2_{AB\_shared} = r^2_A + r^2_B - r^2_{A+B}$ (variance shared by both models)
2) $r^2_{A\_unique} = r^2_{A+B} - r^2_B$ (variance unique to model A)
3) $r^2_{B\_unique} = r^2_{A+B} - r^2_A$ (variance unique to model B)

For three feature spaces:

4) $r^2_{ABC\_shared} = r^2_{A+B+C} + r^2_A + r^2_B + r^2_C - r^2_{A+B} - r^2_{A+C} - r^2_{B+C}$ (variance shared by all three

models)

5) $r^2_{AB\_shared\_noC} = r^2_A + r^2_B - r^2_{A+B} - r^2_{ABC\_shared}$ (variance shared by models A and B, but not C)
6) $r^2_{AC\_shared\_noB} = r^2_A + r^2_C - r^2_{A+C} - r^2_{ABC\_shared}$ (variance shared by models A and C, but not B)
7) $r^2_{BC\_shared\_noA} = r^2_B + r^2_C - r^2_{B+C} - r^2_{ABC\_shared}$ (variance shared by models B and C, but not A)
8) $r^2_{A\_unique} = r^2_{A+B+C} - r^2_{B+C}$ (variance unique to model A)
9) $r^2_{B\_unique} = r^2_{A+B+C} - r^2_{A+C}$ (variance unique to model B)
10) $r^2_{C\_unique} = r^2_{A+B+C} - r^2_{A+B}$ (variance unique to model C)

r^2_A is the variance explained when model A alone is fit; r^2_{A+B} is the variance explained when features from models A and B are jointly fit; and r^2_{A+B+C} is the variance explained when features from models A, B, and C are jointly fit; and so on.

Note that performing variance partitions using the held-out validation set can sometimes result in impossible partitions of variance (i.e. negative variance explained). This can result from joint models having a poorer fit relative to each individual model on its own. In previous work, such estimates were corrected by finding the minimum bias term for which adding this term to the partitions would result in sensible results (i.e. all variance partitions above zero; de Heer et al., 2017). However, an issue with that approach is that such an estimate may *inflate* the significance of results across subjects, since no mean variance value can be below zero after adjustment. Thus, we chose instead not to adjust the partition estimates. Although this results in some impossible values (e.g., the shared variance of Binder and Semantic Structure being significantly negative), these results provide a lower bound on the possible unique variance attributable to each semantic model or sets of models. Thus, if a region shows variance explained significantly *greater* than zero, we can be confident this is indeed the case, but a null result here would be inconclusive.

V. DISCUSSION

Recognizing events is crucial for guiding our social behavior. To return to the Introduction, in Figure 1.1, did you see kissing or biting? Was it the red player who acted on his blue opponent, or the other way around? Although it is conceivable that recognizing such structure in the world requires explicit reasoning about the physical and mental states of the entities involved, in the current thesis, we found evidence to the contrary. The data from our studies suggests instead that the mind automatically extracts structured event information from visual and linguistic input.

In Chapter 2, we found behavioral evidence that one component of event structure – event roles – is extracted from visual scenes. Such extraction was spontaneous and occurred even when event information is irrelevant to the task. By the nature of the task we employed, we can conclude that the computation itself was rapid, occurring within a few hundred milliseconds. Furthermore, the representations of role that were extracted were at an event-general level: participants were not simply encoding actors in a scene as *kicker/kickee*, or *kisser/kissee*, but rather at the more abstract level of *Agent-like* and *Patient-like*. This is precisely what we would expect to observe if the visual system encodes the observed scene in terms of a structured interaction in which one entity, the Agent, performs some act on another, the Patient.

In Chapter 3, we used fMRI to identify brain regions which house representations of action categories invariant to incidental visual properties. Since our hypothesis is that events are defined by relationships between entities, rather than by full dynamic sequences per se, we predicted that such representations should be identifiable even when we decode across multivoxel fMRI patterns elicited by static snapshots (e.g., a photograph of biting) and dynamic sequences. We identified representations with just such properties in brain areas previously observed to respond to action observation. Surprisingly, this included regions previously shown to support recognition motion patterns and groups of interacting objects (section 1.2 in the Introduction), including motion-selective hMT+ and object-selective LOC. This suggests that such regions may code for actions in a manner more abstract than previously thought, perhaps through an experience-driven association between static snapshots of actions and full action sequences. Overall the findings from this study suggest that the representations in these

regions may provide a link between systems which support perceptual recognition and conceptual systems which support complex thought about events. These conceptual systems have not yet been identified in the current studies (see section 5.1 below), but we envision here representations that are non-linguistic in nature and support combinatorial thought (Jackendoff, 2002).

Finally, in Chapter 4, we returned to language, which was the initial inspiration for our investigation of event structure in vision. We sought to provide support for the hypothesis that the brain encodes the meanings of verbs in part by the structure of the event that they refer to. We used a voxel-wise encoding model approach to test several predictions of the relationship between linguistic and semantic structure. These predictions were confirmed. We also compared a model of event structure alongside other leading models in the field (a distributional semantic model and sensorimotor/ cognitive model), finding that all three models shared substantial explained variance in language-selective regions. The work in this chapter suggests that properties of an event's semantic structure (e.g. Agent, Cause, State) are encoded spontaneously by people as they comprehend naturalistic speech, and that such properties are implicitly present in the representations of other semantic models.

Together, these studies provide support for the hypothesis that perception itself traffics in event structure, of the kind predicted by patterns observed in the linguistic structures associated with verbs. These studies also identify candidates for areas of the brain that may support abstract representations of events, from both visual and linguistic input. More generally, our work demonstrates the utility of using language as a window into the high-level representations that may be afforded by the visual system.

5.1. Limitations and Future Directions

In Chapter 2, we constrained our investigation of the perception of event structure to a core distinction between Agent and Patient, which is consistent with theories that posit such a coarse distinction in language (Dowty, 1991). However, a stronger test of this coarse distinction would require investigation of a larger set of distinctions made in the theoretical domain of event roles in language. In particular, some theories posit additional roles to account for how we understand differences between elements that occupy the same syntactic position (Levin & Rappaport-Hovav, 2005); for example,

between a Recipient and Destination in *I gave the man [Recipient] a book*, vs. *I loaded the wagon [Destination] with hay*. Conducting our task with additional roles would allow us to understand precisely what event role distinctions the mind makes.

We also note that in Chapter 3, we did not identify neural representations of event role identity elicited by visual input (the entity bound to its corresponding role, i.e. whether actor A was the Agent and B the Patient or vice-versa, not simply the presence of an Agent or Patient alone). Given that we know these representations must exist (Chapter 2), there are several possibilities for this. One of course is that we might not have had sufficient power to detect these representations using fMRI. This is not surprising: previous work identifying such representations in the linguistic domain required several tens of subjects to observe such an effect (Frankland & Greene, 2015), while we only had 15. Additionally, the coding of role identity may simply be too sparse to be detectable using the multivariate fMRI techniques we employed. Indeed, decoding representations of person identity is notoriously fickle (Anzellotti & Caramazza, 2014). Techniques such as fMRI adaptation can assist in these cases. Identifying such representations could be a key element of understanding what brain regions are involved in building structured event representations from a visual scene (Chapter 2).

In Chapter 3, we found evidence for neural representations of action categories that generalized across incidental perceptual properties. However, the data thus far does not speak to the organizational principles underlying representation of events in these regions. Does a model of event structure derived from the linguistic domain of the kind used in Chapter 4 predict the representation of events from visual input in these regions? We were limited by our relatively small set of categories here, so future work should widely sample the rich space of events to model such representations.

Relatedly, an open question is how information about event structure in language and vision are integrated. There is good reason to suspect such common conceptual representations exist (see section 1.1 in Introduction). However, we have not yet identified a common locus for event structure across modalities, as our investigations in vision and language proceeded in parallel. Based on the results of Chapters 3 and 4, we speculate that the posterior middle temporal cortex (pMTC) may play a role in integrating event information from the two modalities (Lingnau & Downing, 2015). Despite anatomical differences across the two studies, they both showed strong results in

adjacent regions of pMTC, with language results anterior to the visual (Figures 3.2 and 4.6). The view that pMTC may support such integration is bolstered by a recent study showing common representation of actions across language and vision (M. F. Wurm & Caramazza, 2018), although explicit models of event structure have not yet been tested here. In our approach (Chapter 4), this could be investigated by training encoding models of event structure on linguistic stimuli and testing them in a held-out validation set of visual events (e.g. in short films), or vice-versa.

A still unresolved question is why there should be a correspondence between event structure and patterns of linguistic structure in the first place. Recall in the Introduction and in Chapter 4 that properties of event structure are predictive of the sets of linguistic structures a verb can appear in. For example, the number of noun phrases can be predicted by the nature of the event: *sleep* can appear in sentences like *Jimmy sleeps,* but sound odd or ungrammatical in sentences like *Jimmy sleeps his ice cream,* precisely because sleeping involves one entity performing the action. In Chapter 4, we demonstrated the presence of this mapping between linguistic structure and event structure in the fully linguistic adult brain, such that hearing a verb activates both its associated semantic structure and its associated syntactic frames. However, this does not explain how the association came to be, both in the brain, and in general through the evolutionary course of language.

Several proposals have been proffered to explain this correspondence. Strickland (2016) proposes that "core" knowledge available to the infant early on – possibly in perception – makes certain grammatical distinctions more salient (e.g., syntactic structures correlated with Causation), such that elements not associated with core cognition (e.g., the distinction between tables and chairs) are less likely to get coded in human grammatical systems as languages evolve. This dovetails nicely with Pinker's (1989) proposal of how children acquire the syntactic structures of their language, called semantic bootstrapping. Given that the space of possible grammars for language is immense, Pinker proposes that children use event structure gleaned from observation, along with unlearned mappings between semantics and syntax (e.g. objects map to nouns), to acquire their language's grammar. Crucially, Pinker's proposal requires that high-level representations of events are available from observation in the first place, before the child has access to the structural principles of their grammar. Our work in

Chapter 2 supports this proposal suggesting that at least in the adult visual system, aspects of event structure (here, roles) come "for free", and our ongoing work suggests this is also the case in young preschool children (not published). An alternative theory is that an association between semantics and syntax is necessary for the child to learn the full range of meanings to which verbs and other words refer (Fisher et al., 1991; Gleitman, 1990; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). After all, what perceptual input would a child use to learn what *thinking* means, given that it has no overt perceptual correlates? Gleitman and colleagues argue for an inverse procedure to Pinker's, whereby children use the syntactic frames of a verb to constrain and predict its meaning (what they call syntactic bootstrapping). Importantly, all of these proposals share the requirement that the mind have some kind of non-linguistic conceptual structure, capable of interfacing with both linguistic and perceptual systems (Jackendoff, 2002). Thus, identifying such non-linguistic conceptual structure is an important target of future investigations.

One important path for future work is identifying the computations that give rise to high-level event structure representations from visual input. We have identified the perceptual phenomenon of study (Chapter 2) and some of the cortical regions that house representations at least partially abstracted from the visual input (Chapter 3). However, we have little to say about how these representations "got there" in the first place, or about their role in other cognitive processes such as memory or language. A class of recent models known as hierarchical convolutional neural networks (HCNN) offers the potential for breakthroughs in the computational understanding of high-level visual abstraction for recognition. HCNNs have recently achieved remarkable performance at tasks like object, scene, and action recognition (Simonyan & Zisserman, 2014) and are the current leading computational models of information processing in the human visual system (Bonner & Epstein, 2018; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). They achieve this remarkable feat in part by capitalizing on statistical patterns in natural scene input. For example, the presence of a person and a hoagie in a scene may be a reliable heuristic for *person eating sandwich*.

However, we suspect that to recognize situations as flexibly as the human mind does, such statistical heuristics alone cannot be used for event understanding, especially when novel entities are involved in novel situations. Consider the ease with which you

recognized the biting scene in Figure 1.1, despite never having observed such a scene before (Lake, Ullman, Tenenbaum, & Gershman, 2016). We hypothesize that since events capture not only atomic entities, but relationships between entities (as we demonstrated in Chapter 2), statistical patterns alone may not be sufficient. One important component that they may be missing is the principle of compositionality, akin to the similar principle in linguistics and semantics: that entities in an event representation go together in structured, lawful ways (George et al., 2017; Yuille & Liu, 2018). In future work, we can combine current approaches in mathematical modeling of high-level visual tasks, representational similarity analyses between models and the brain (Kriegeskorte et al., 2008), and techniques to probe the internal representations of such models (Bonner & Epstein, 2018) to provide insights into the possible algorithms and computational architectures that may underlie the recognition of events.

5.2. Conclusions

In this thesis we have argued that the mind extracts an event structure from the observed world whose representational content can be predicted by patterns of linguistic structure (with elements such as Cause, Motion, State Change, and event roles). We have contributed to an understanding of how the mind works at several levels (Marr, 1982), by identifying a key computational-level problem to solve (understanding the structure of events), its domain (perception and language; Chapters 2 and 4), and candidate brain areas where the representations for event structure may be (Chapters 3 and 4). Thus, our studies lay the groundwork for future investigation of these issues.

Taken together, the findings from this thesis suggest that a fundamental process of the mind is analyzing the structure of who is doing what to whom. We may literally *perceive* a red-shirted soccer player biting his opponent, even if we ultimately never learn why.

APPENDICES

Appendix A

Examples of images used in Chapter 2

An example image for each event category featured in the experiments (for Experiments 1a and 1b, Female Agent on the Left images; for Experiments 2 and 3, Blue Agent on the Left images). Agent and Patient poses were similar for the four versions of each event category. Although the images used in Experiments 2 and 3 were desaturated to a level of 3% to make the task (color search) more difficult, they are shown here in full color for illustrative purposes. See sections 2.1.2, 4.1.2, and 5.1.2 of Chapter 2 for details.

Experiments 1a and 1b (Gender Search)

brush chase feed film kick

look at punch push scratch tap

Experiment 2 (Color Search)

bandage kick kiss poke pull

scratch slap stab strangle tickle

Experiment 3 (Color Search, Mirror-Flipped)

bandage kick kiss poke pull

scratch slap stab strangle tickle

BIBLIOGRAPHY

Abdollahi, R. O., Jastorff, J., & Orban, G. A. (2013). Common and segregated processing of observed actions in human SPL. Cerebral Cortex, 23(11), 2734–2753. http://doi.org/10.1093/cercor/bhs264

Abend, O., Reichart, R., & Rappoport, A. (2008). A supervised algorithm for verb disambiguation into VerbNet classes. Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08, (August), 9–16. http://doi.org/10.3115/1599081.1599083

Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. Psychological Science, 14(5), 427–432. http://doi.org/10.1111/1467-9280.01458

Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. NeuroImage, 35(4), 1480–94. http://doi.org/10.1016/j.neuroimage.2007.02.005

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., … Raizada, R. D. S. (2017). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cerebral Cortex, 27(9), 4379–4395. http://doi.org/10.1093/cercor/bhw240

Anderson, A. J., Lalor, E. C., Lin, F., Binder, J. R., Fernandino, L., Humphries, C. J., … Wang, X. (2018). Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. Cerebral Cortex, (May), 1–16. http://doi.org/10.1093/cercor/bhy110

Anzellotti, S., & Caramazza, A. (2014). The neural mechanisms for the recognition of face identity in humans. Frontiers in Psychology, 5(June), 672. http://doi.org/10.3389/fpsyg.2014.00672

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. International Journal of Psychology Research, 3, 12–28. http://doi.org/10.21500/20112084.80

Bach, P., Peelen, M. V, & Tipper, S. P. (2010). On the role of object information in action observation: an fMRI study. Cerebral Cortex (New York, N.Y. : 1991), 20(12), 2798–809. http://doi.org/10.1093/cercor/bhq026

Baillargeon, R., Stavans, M., Wu, D., Gertner, Y., Setoh, P., Kittredge, A. K., & Bernard, A. (2012). Object individuation and physical reasoning in infancy: An integrative account. Language Learning and Development, 8(1), 4–46. http://doi.org/10.1080/15475441.2012.630610

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2016). Human-object interactions are more than the sum of their parts. Cerebral Cortex, 1–13. http://doi.org/10.1093/cercor/bhw077

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., Norman, K. A., … Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. Neuron, 95(3), 709–721.e5. http://doi.org/10.1016/j.neuron.2017.06.041

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: the influence of trial history and data transformations. Journal of Experimental Psychology. Learning, Memory, and Cognition, 39(5), 1563–71. http://doi.org/10.1037/a0032186

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In EMNLP-2010 (pp. 1183–1193). http://doi.org/10.4249/scholarpedia.3881

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Green, P. (2016). lme4: Linear Mixed-Effects Models using “Eigen” and S4 (Version 1.1-12). Retrieved November 1, 2016, from https://cran.r-project.org/package=lme4

Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A., & Saxe, R. (2008). Concepts are more than percepts: the case of action verbs. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 28(44), 11347–53. http://doi.org/10.1523/JNEUROSCI.3039-08.2008

Bedny, M., Caramazza, A., Pascual-Leone, A., & Saxe, R. (2012). Typical neural

representations of action verbs develop without vision. Cerebral Cortex (New York, N.Y. : 1991), 22(2), 286–93. http://doi.org/10.1093/cercor/bhr081

Bedny, M., Dravida, S., & Saxe, R. (2014). Shindigs, brunches, and rodeos: The neural basis of event words. Cognitive, Affective & Behavioral Neuroscience, 14(3), 891–901. http://doi.org/10.3758/s13415-013-0217-z

Biederman, I., Blickle, T. W., Teitelbaum, R. C., & Klatsky, G. J. (1988). Object search in nonscene displays. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 456–467. http://doi.org/10.1037/0278-7393.14.3.456

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. Cognitive Psychology, 14(2), 143–77.

Binder, J. R. (2016). In defense of abstract conceptual representations. Psychonomic Bulletin & Review, 23, 1096–1108. http://doi.org/10.3758/s13423-015-0909-1

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. Cognitive Neuropsychology, 33(3–4), 130–174. http://doi.org/10.1080/02643294.2016.1147426

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 (New York, N.Y. : 1991), 19(12), 2767–96. http://doi.org/10.1093/cercor/bhp055

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. NeuroImage, 127, 307–323. http://doi.org/10.1016/j.neuroimage.2015.11.069

Blank, I., & Fedorenko, E. (2017). Domain-general brain regions do not track linguistic input as closely as language-selective regions. The Journal of Neuroscience, 3642–16. http://doi.org/10.1523/JNEUROSCI.3642-16.2017

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLoS Computational Biology

(Vol. 14). http://doi.org/10.1371/journal.pcbi.1006111

Bracci, S., & Peelen, M. V. (2013). Body and object effectors: the organization of object representations in high-level visual cortex reflects body-object interactions. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33(46), 18247–58. http://doi.org/10.1523/JNEUROSCI.1322-13.2013

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433–436. http://doi.org/10.1163/156856897X00357

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. Brain and Language, 120(2), 163–173. http://doi.org/10.1016/j.bandl.2010.04.002

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W. M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. Brain and Language, 157–158, 81–94. http://doi.org/10.1016/j.bandl.2016.04.008

Bressler, D. W., & Silver, M. A. (2010). Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. NeuroImage, 53(2), 526–533. http://doi.org/10.1016/j.neuroimage.2010.06.063

Brown, P. M., & Dell, G. S. (1987). Adapting Production to Comprehension: The Explicit Mention of Instruments. Cognitive Psychology, 19, 441–472.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46(3), 904–911. http://doi.org/10.3758/s13428-013-0403-5

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., … Rizzolatti, G. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. Journal of Cognitive Neuroscience, 16(1), 114–26. http://doi.org/10.1162/089892904322755601

Caramazza, A., Anzellotti, S., Strnad, L., & Lingnau, A. (2014). Embodied Cognition and Mirror Neurons: A Critical Assessment. Annual Review of Neuroscience, 37, 1–15.

http://doi.org/10.1146/annurev-neuro-071013-013950

Carey, S. (2009). The Origin of Concepts. The Origin of Concepts, (March), 1–608. http://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. NeuroImage, 50(3), 1148–67. http://doi.org/10.1016/j.neuroimage.2009.12.112

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. Journal of Experimental Psychology. Human Perception and Performance, 33(4), 753–63. http://doi.org/10.1037/0096-1523.33.4.753

Cattaneo, L., Sandrini, M., & Schwarzbach, J. (2010). State-dependent TMS reveals a hierarchical representation of observed acts in the temporal, parietal, and premotor cortices. Cerebral Cortex (New York, N.Y. : 1991), 20(9), 2252–8. http://doi.org/10.1093/cercor/bhp291

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2016). Shared memories reveal shared structure in neural activity across individuals. Nat Neurosci, advance on(1). http://doi.org/10.1038/nn.4450

Chen, L., & Eugenio, B. Di. (2010). A Maximum Entropy Approach To Disambiguating VerbNet Classes.

Cheries, E. W., Wynn, K., & Scholl, B. J. (2006). Interrupting infants' persisting object representations: An object-based limit? Developmental Science, 9(5), 50–58. http://doi.org/10.1111/j.1467-7687.2006.00521.x

Chong, T. T.-J., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. Current Biology : CB, 18(20), 1576–80. http://doi.org/10.1016/j.cub.2008.08.068

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: the processing of semantic roles in visual narrative. Cognitive Psychology, 67(3), 73–97. http://doi.org/10.1016/j.cogpsych.2013.07.002

Cohn, N., Paczynski, M., & Kutas, M. (2017). Not so secret agents: Event-related potentials to semantic roles in visual event comprehension. Brain and Cognition, 119(April), 1–9. http://doi.org/10.1016/j.bandc.2017.09.001

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 : The Official Journal of the Society for Neuroscience, 32(8), 2608–18. http://doi.org/10.1523/JNEUROSCI.5547-11.2012

Connor, D. H. O., Fukui, M. M., Pinsk, M. A., Kastner, S., O'Connor, D. H., Fukui, M. M., … Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. Nature Neuroscience, 5(11), 1203–1209. http://doi.org/10.1038/nn957

Croft, W. (2012). Verbs: Aspect and Causal Structure. Oxford: Oxford University Press.

De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. Cognition, 178(November 2017), 133–146. http://doi.org/10.1016/j.cognition.2018.05.017

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. The Journal of Neuroscience, 37(27), 3267–16. http://doi.org/10.1523/JNEUROSCI.3267-16.2017

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. Cerebral Cortex (New York, N.Y. : 1991), 1–14. http://doi.org/10.1093/cercor/bhv111

Dinstein, I., Gardner, J. L., Jazayeri, M., & Heeger, D. J. (2008). Executed and observed movements have different distributed representations in human aIPS. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 28(44), 11231–9. http://doi.org/10.1523/JNEUROSCI.3585-08.2008

Dobel, C., Diesendruck, G., & Bölte, J. (2007). How writing system and age influence spatial representations of actions: a developmental, cross-linguistic study. Psychological Science, 18(6), 487–91. http://doi.org/10.1111/j.1467-9280.2007.01926.x

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly

seen. Acta Psychologica, 125(2), 129–43. http://doi.org/10.1016/j.actpsy.2006.07.004

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science (New York, N.Y.), 293(5539), 2470–3. http://doi.org/10.1126/science.1063414

Dowty, D. (1991). Thematic proto-roles and argument selection. Language, 67(3), 547–619.

Dryer, M. S. (2013). Order of Subject, Object and Verb. In M. S. Dryer & M. Haspelmath (Eds.), The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Epstein, R. A., & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598–601. http://doi.org/10.1038/33402

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nature Methods, 16(1), 111–116. http://doi.org/10.1038/s41592-018-0235-4

Etzel, J. A., Gazzola, V., & Keysers, C. (2008). Testing simulation theory with cross-modal multivariate classification of fMRI data. PloS One, 3(11), e3690. http://doi.org/10.1371/journal.pone.0003690

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33(25), 10552–8. http://doi.org/10.1523/JNEUROSCI.0051-13.2013

Fausey, C. M., Long, B. L., Inamori, A., & Boroditsky, L. (2010). Constructing agency: the role of language. Frontiers in Psychology, 1(October), 162. http://doi.org/10.3389/fpsyg.2010.00162

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. http://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Nieto-castañón, A., & Kanwisher, N. (2013). Syntactic processing in the human brain: What we know, what we don't know, and a suggestion for how to proceed, 120(2), 187–207. http://doi.org/10.1016/j.bandl.2011.01.001.Syntactic

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. Proceedings of the National Academy of Sciences, 201612132. http://doi.org/10.1073/pnas.1612132113

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. Annals of the New York Academy of Sciences, 1369(1), 132–153. http://doi.org/10.1111/nyas.13046

Feldman, H., Goldin-Meadow, S., & Gleitman, L. R. (1978). Beyond Herodotus: The creation of language by linguistically deprived deaf children. In A. Lock (Ed.), Action, Symbol, and Gesture: The Emergence of Language (pp. 351–414). New York: Academic Press.

Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S., & Binder, J. R. (2016). Heteromodal Cortical Areas Encode Sensory-Motor Features of Word Meaning. Journal of Neuroscience, 36(38), 9763–9769. http://doi.org/10.1523/JNEUROSCI.4095-15.2016

Fernandino, L., Humphries, C. J., Seidenberg, M. S., Gross, W. L., Conant, L. L., & Binder, J. R. (2015). Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. Neuropsychologia, 76, 17–26. http://doi.org/10.1016/j.neuropsychologia.2015.04.009

Ferri, S., Kolster, H., Jastorff, J., & Orban, G. A. (2013). The overlap of the EBA and the MT/V5 cluster. NeuroImage, 66, 412–425. http://doi.org/10.1016/j.neuroimage.2012.10.060

Ferri, S., Rizzolatti, G., & Orban, G. A. (2015). The organization of the posterior parietal cortex devoted to upper limb actions: An fMRI study. Human Brain Mapping, 36(10), 3845–3866. http://doi.org/10.1002/hbm.22882

Fillmore, C. J. (1968). The Case for Case. Texas Symposium on Language Universals.

http://doi.org/10.2307/326399

Firestone, C., & Keil, F. C. (2016). Seeing the Tipping Point: Balance Perception and Visual Shape. Journal of Experimental Psychology: General, 145(7), 872–881. http://doi.org/10.1037/xge0000151

Fischl, B., Sereno, M. I., Tootell, R. B. H., & Dale, A. M. (1999). High-resolution inter-subject averaging and a surface-based coordinate system. Human Brain Mapping, 8(FEBRUARY 1999), 272–284. http://doi.org/10.1002/(SICI)1097-0193(1999)8

Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. Cognitive Psychology, 23(3), 331–392. http://doi.org/10.1016/0010-0285(91)90013-E

Fleischer, F., Caggiano, V., Thier, P., & Giese, M. a. (2013). Physiologically inspired model for the visual recognition of transitive hand actions. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33(15), 6563–80. http://doi.org/10.1523/JNEUROSCI.4129-12.2013

Fodor, J. A. (1983). The Modularity of Mind. Boston: MIT Press.

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. Proceedings of the National Academy of Sciences of the United States of America, 112(37), 11732–11737. http://doi.org/10.1073/pnas.1421236112

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. Perception & Psychophysics, 33(6), 575–581. http://doi.org/10.3758/BF03202940

Fyshe, A., Talukdar, P., Murphy, B., & Mitchell, T. (2014). Interpretable Semantic Vectors from a Joint Model of Brain-and Text-Based Meaning. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 489–499. http://doi.org/10.14440/jbm.2015.54.A

Fyshe, A., Wehbe, L., Talukdar, P., Murphy, B., & Mitchell, T. (2015). A compositional and interpretable semantic space. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human

Language Technologies (NAACL HLT 2015), 32–41.

Gallivan, J. P., Adam McLean, D., Valyear, K. F., & Culham, J. C. (2013). Decoding the neural mechanisms of human tool use. ELife, 2013(2), 1–29. http://doi.org/10.7554/eLife.00425

Gallivan, J. P., Chapman, C. S., Mclean, D. A., Flanagan, J. R., & Culham, J. C. (2013). Activity patterns in the category-selective occipitotemporal cortex predict upcoming motor actions. European Journal of Neuroscience, 38(3), 2408–2424. http://doi.org/10.1111/ejn.12215

Gallivan, J. P., & Culham, J. C. (2015). Neural coding within human brain areas involved in actions. Current Opinion in Neurobiology, 33, 141–149. http://doi.org/10.1016/j.conb.2015.03.012

Gallivan, J. P., McLean, D. A., Smith, F. W., & Culham, J. C. (2011). Decoding effector-dependent and effector-independent movement intentions from human parieto-frontal brain activity. Journal of Neuroscience, 31(47), 17149–17168. http://doi.org/10.1523/JNEUROSCI.1058-11.2011

Gao, T., Scholl, B. J., & McCarthy, G. (2012). Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 32(41), 14276–80. http://doi.org/10.1523/JNEUROSCI.0562-12.2012

Gazzola, V., van der Worp, H., Mulder, T., Wicker, B., Rizzolatti, G., & Keysers, C. (2007). Aplasics Born without Hands Mirror the Goal of Hand Actions with Their Feet. Current Biology, 17(14), 1235–1240. http://doi.org/10.1016/j.cub.2007.06.045

George, D., Lehrach, W., Kansky, K., Lazaro-Gredilla, M., Laan, C., Marthi, B., … Liu, Y. (2017). A Generative Vision Model that Trains with High Data Efficiency. Science, 10(October), 1–19. http://doi.org/10.1126/science.aag2612

Gervais, W. M., Reed, C. L., Beall, P. M., & Roberts, R. J. (2010). Implied body action directs spatial attention. Attention, Perception & Psychophysics, 72(6), 1437–43. http://doi.org/10.3758/APP.72.6.1437

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A Large-

Scale Evaluation Set of Verb Similarity.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience, 4(3), 179–92. http://doi.org/10.1038/nrn1057

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. Psychonomic Bulletin & Review. http://doi.org/10.3758/s13423-016-1004-y

Gleitman, L. R. (1990). The structural sources of verb meanings. Language Acquisition, 1(1), 3–55. http://doi.org/10.1207/s15327817la0101_2

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. Language Learning and Development, 1(1), 23–64.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. Journal of Memory and Language, 57(4), 544–569. http://doi.org/10.1016/j.jml.2007.01.007

Goldberg, A. E. (1999). The Emergence of the Semantics of Argument Structure Constructions. In B. MacWhinney (Ed.), The Emergence of Language (pp. 197–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldin-Meadow, S., & Feldman, H. (1977). The Development of Language-Like Communication Without a Language Model. Science, 197(4301), 401–403.

Goldin-Meadow, S., So, W. C., Ozyürek, A., & Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. Proceedings of the National Academy of Sciences of the United States of America, 105(27), 9163–8. http://doi.org/10.1073/pnas.0710060105

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data, 3, 160044. http://doi.org/10.1038/sdata.2016.44

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually

grouped. Journal of Experimental Psychology. Human Perception and Performance, 32(5), 1107–19. http://doi.org/10.1037/0096-1523.32.5.1107

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. Journal of Vision, 14(1), 1–11. http://doi.org/10.1167/14.1.14.doi

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognitive Psychology, 58(2), 137–76. http://doi.org/10.1016/j.cogpsych.2008.06.001

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nature Reviews. Neuroscience, 15, 536–548. http://doi.org/10.1038/nrn3747

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. Neuron, 35(6), 1167–75.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. Journal of Cognitive Neuroscience, 12(5), 711–20.

Gruber, J. S. (1965). Studies in lexical relations. In PhD Dissertation. Cambridge, MA.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. Journal of Experimental Psychology: General, 142(3), 880–905. http://doi.org/10.1037/a0030045

Hafri, A., Trueswell, J. C., & Epstein, R. A. (2017). Neural representations of observed actions generalize across static and dynamic visual input. The Journal of Neuroscience, 37(11), 2496–16. http://doi.org/10.1523/JNEUROSCI.2496-16.2017

Hafri, A., Trueswell, J. C., & Strickland, B. (2016). Extraction of event roles from visual scenes is rapid, automatic, and interacts with higher-level visual processing. In Proceedings of the 38th Annual Conference of the Cognitive Science Society. Philadelphia, PA.

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual

scenes is rapid, spontaneous, and interacts with higher-level visual processing. Cognition, 175, 36–52. http://doi.org/https://doi.org/10.1016/j.cognition.2018.02.011

Hamilton, A. F. D. C., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 26(4), 1133–7. http://doi.org/10.1523/JNEUROSCI.4551-05.2006

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. Nature, 450(7169), 557–9. http://doi.org/10.1038/nature06288

Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Meaningful differences in the everyday experience of young American children. Baltimore, MD, US: Paul H Brookes Publishing.

Hartshorne, J. K. (2014). What is implicit causality? Language, Cognition and Neuroscience, 29(7), 804–824. http://doi.org/10.1080/01690965.2013.796396

Hartshorne, J. K., Bonial, C., & Palmer, M. (2014). The VerbCorner Project : Findings from Phase 1 of Crowd-Sourcing a Semantic Decomposition of Verbs. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, (1989), 397–402.

Haxby, J., Gobbini, M., Furey, M., & Ishai, A. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(September), 2425–30.

Hernández, M., Fairhall, S. L., Lenci, A., Baroni, M., & Caramazza, A. (2014). Predication Drives Verb Cortical Signatures. Journal of Cognitive Neuroscience. http://doi.org/10.1162/jocn

Hickok, G. (2009). Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. Journal of Cognitive Neuroscience, 21(7), 1229–1243. http://doi.org/10.1162/jocn.2009.21189

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews. Neuroscience, 8(5), 393–402. http://doi.org/10.1038/nrn2113

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70. http://doi.org/10.2307/4615733

Hume, D. (1739). A Treatise of Human Nature. Oxford: Clarendon Press.

Huth, A. G., Heer, W. A. De, Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453–458. http://doi.org/10.1038/nature17637

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. Proceedings of the National Academy of Sciences, 201714471. http://doi.org/10.1073/pnas.1714471114

Jackendoff, R. S. (1990). Semantic Structures. Cambridge, MA: MIT Press. http://doi.org/10.1037/031829

Jackendoff, R. S. (2002). Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press.

Jain, S., & Huth, A. (2018). Incorporating Context into Language Encoding Models for fMRI. BioRxiv, 327601. http://doi.org/10.1101/327601

Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., & Orban, G. A. (2010a). Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. Journal of Neurophysiology, 104(1), 128–140. http://doi.org/10.1152/jn.00254.2010

Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., & Orban, G. a. (2010b). Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. Journal of Neurophysiology, 104(1), 128–140. http://doi.org/10.1152/jn.00254.2010

Jastorff, J., Kourtzi, Z., & Giese, M. A. (2009). Visual learning shapes the processing of complex movement stimuli in the human brain. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 29(44), 14026–38. http://doi.org/10.1523/JNEUROSCI.3070-09.2009

Jellema, T., & Perrett, D. I. (2006). Neural representations of perceived bodily actions

using a categorical frame of reference. Neuropsychologia, 44(9), 1535–46. http://doi.org/10.1016/j.neuropsychologia.2006.01.020

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage, 17(2), 825–841. http://doi.org/10.1016/S1053-8119(02)91132-8

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. NeuroImage, 60(4), 2357–64. http://doi.org/10.1016/j.neuroimage.2012.02.055

Kable, J. W., & Chatterjee, A. (2006). Specificity of action representations in the lateral occipitotemporal cortex. Journal of Cognitive Neuroscience, 18(9), 1498–517. http://doi.org/10.1162/jocn.2006.18.9.1498

Kako, E. (2006). Thematic role properties of subjects and objects. Cognition, 101(1), 1–42. http://doi.org/10.1016/j.cognition.2005.08.002

Kann, K., Warstadt, A., Williams, A., & Bowman, S. R. (2019). Verb Argument Structure Alternations in Word and Sentence Embeddings. Proceedings Ofthe Society for Computation in Linguistics (SCiL) 2019, 287–297.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the United States of America, 107(25), 11163–11170. http://doi.org/10.1073/pnas.1005062107

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 17(11), 4302–11. http://doi.org/10.1098/Rstb.2006.1934

Kartsaklis, D. (2014). Compositional Operators in Distributional Semantics. Springer Science Reviews, 2(1–2), 161–177.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(3), 352–355. http://doi.org/10.1038/nature06713

Kemmerer, D., & Gonzalez-Castillo, J. (2010). The Two-Level Theory of verb meaning: An approach to integrating the semantics of action with the mirror neuron system. Brain and Language, 112(1), 54–76. http://doi.org/10.1016/j.bandl.2008.09.010

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Computational Biology, 10(11), e1003915. http://doi.org/10.1371/journal.pcbi.1003915

Kilner, J. M. (2011). More than one pathway to action understanding. Trends in Cognitive Sciences, 15(8), 352–7. http://doi.org/10.1016/j.tics.2011.06.005

Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 29(32), 10153–9. http://doi.org/10.1523/JNEUROSCI.2668-09.2009

Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? Cerebral Cortex (New York, N.Y. : 1991), 21(8), 1738–46. http://doi.org/10.1093/cercor/bhq240

Kim, J. G., Biederman, I., & Juan, C.-H. (2011). The benefit of object interactions arises in the lateral occipital cortex independent of attentional modulation from the intraparietal sulcus: a transcranial magnetic stimulation study. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 31(22), 8320–4. http://doi.org/10.1523/JNEUROSCI.6450-10.2011

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. Language Resources and Evaluation, 42(1), 21–40. http://doi.org/10.1007/s10579-007-9048-2

Kline, M., Muentener, P., & Schulz, L. (2013). Transitive and periphrastic sentences affect memory for simple causal scenes, 1, 1–5.

Kocagoncu, E., Clarke, A., Devereux, B. J., & Tyler, L. K. (2017). Decoding the Cortical Dynamics of Sound-Meaning Mapping. The Journal of Neuroscience, 37(5), 1312–1319. http://doi.org/10.1523/JNEUROSCI.2858-16.2016

Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., & Keil, F. C. (2017). Categories and Constraints in Causal Perception. Psychological Science.

http://doi.org/10.1177/0956797617719930

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. Journal of Cognitive Neuroscience, 12(1), 48–55.

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. Science (New York, N.Y.), 293(5534), 1506–9. http://doi.org/10.1126/science.1061133

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., Maciver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron, 93(3), 480–490. http://doi.org/10.1016/j.neuron.2016.12.041

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), 3863–8.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. Trends in Cognitive Sciences, 17(8), 401–12. http://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2(November), 4. http://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience, 12(5), 535–40. http://doi.org/10.1038/nn.2303

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. Psychological Science, 14(5), 402–8.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. Behavioral and Brain Sciences, 1–101. http://doi.org/10.1017/S0140525X16001837

Landau, B., & Gleitman, L. R. (1985). Language and Experience: Evidence from the Blind Child. Cambridge, MA, MA: Harvard University Press.

Langacker, R. (1987). Foundations of Cognitive Grammar. Stanford: Stanford University Press.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 26(11), 2894–906. http://doi.org/10.1523/JNEUROSCI.4915-05.2006

Lenci, A. (2018). Distributional Models of Word Meaning. Annual Review of Linguistics, (December 2017). http://doi.org/10.1146/annurev-linguistics-030514-125254

Lescroart, M. D., & Gallant, J. L. (2018). Human scene-selective areas represent 3D configurations of surfaces. Neuron.

Leshinskaya, A., & Caramazza, A. (2015). Abstract categories of functions in anterior parietal lobe. Neuropsychologia, 76(2015), 27–40. http://doi.org/10.1016/j.neuropsychologia.2015.01.014

Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. Cognition, 50(1–3), 211–238. http://doi.org/10.1016/0010-0277(94)90029-9

Leslie, A. M. (1995). A Theory of Agency.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? Cognition, 25(April 1986), 265–288.

Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation. Chicago, IL: University of Chicago Press.

Levin, B., & Rappaport-Hovav, M. (2005). Argument Realization. Cambridge, UK: Cambridge University Press.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. Transactions of the Association for Computational Linguistics, 3, 211–225. http://doi.org/10.1186/1472-6947-15-S2-S2

Lingnau, A., & Downing, P. E. (2015). The lateral occipitotemporal cortex in action. Trends in Cognitive Sciences, 19(5), 268–277. http://doi.org/10.1016/j.tics.2015.03.006

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies, 4(1990), 521–535. http://doi.org/10.1148/radiol.2513081056

Mahon, B. Z., Milleville, S. C., Negri, G. A. L., Rumiati, R. I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. Neuron, 55(3), 507–520. http://doi.org/10.1016/j.neuron.2007.07.011

Majdandžić, J., Bekkering, H., Van Schie, H. T., & Toni, I. (2009). Movement-specific repetition suppression in ventral and dorsal premotor cortex during action observation. Cerebral Cortex, 19(11), 2736–2745. http://doi.org/10.1093/cercor/bhp049

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. Journal of Memory and Language, 92, 57–78. http://doi.org/10.1016/j.jml.2016.04.001

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York, NY, USA: Henry Holt and Co., Inc.

Mayrhofer, R., & Waldmann, M. R. (2014). Indicators of causal agency in physical interactions: The role of the prior context. Cognition, 132(3), 485–490. http://doi.org/10.1016/j.cognition.2014.05.013

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs.CL], 1–12. http://doi.org/10.1162/153244303322533223

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. Cognitive Science, 34(8), 1388–1429. http://doi.org/10.1111/j.1551-6709.2010.01106.x

Mouchetant-Rostaing, Y., Giard, M.-H., Bentin, S., Aguera, P.-E., & Pernier, J. (2000). Neurophysiological correlates of face gender processing in humans. European Journal of Neuroscience, 12(1), 303–310. http://doi.org/10.1046/j.1460-9568.2000.00888.x

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. Cognitive Psychology, 61(2), 63–86. http://doi.org/10.1016/j.cogpsych.2010.02.001

Naselaris, T., & Kay, K. N. (2015). Resolving Ambiguities of MVPA Using Explicit Models of Representation. Trends in Cognitive Sciences, 19(10), 551–554. http://doi.org/10.1016/j.tics.2015.07.005

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. Neuron, 63(6), 902–915. http://doi.org/10.1016/j.neuron.2009.09.006

Nichols, T. E., Brett, M., Andersson, J., Poline, J. B., & Wager, T. (2005). Valid conjunction inference with the minimum statistic. NeuroImage, 25(3), 653–60. http://doi.org/10.1016/j.neuroimage.2004.12.005

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples. Human Brain Mapping, 15(1), 1–25. http://doi.org/10.1002/hbm.1058

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology : CB, 21(19), 1641–6. http://doi.org/10.1016/j.cub.2011.08.031

Nonyane, B. A. S., & Theobald, C. M. (2007). Design sequences for sensory studies: achieving balance for carry-over and position effects. The British Journal of Mathematical and Statistical Psychology, 60(Pt 2), 339–349. http://doi.org/10.1348/000711006X114568

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2018). Voxelwise encoding models with non-spherical multivariate normal priors. BioRxiv. http://doi.org/10.1101/386318

Ogawa, K., & Inui, T. (2011). Neural representation of observed actions in the parietal and premotor cortex. NeuroImage, 56(2), 728–735. http://doi.org/10.1016/j.neuroimage.2010.10.043

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision,

42(3), 145–175. http://doi.org/10.1023/A:1011139631724

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2012a). Viewpoint (in)dependence of action representations: an MVPA study. Journal of Cognitive Neuroscience, 24(4), 975–89. http://doi.org/10.1162/jocn_a_00195

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2012b). Visuo-motor imagery of specific manual actions: A multi-variate pattern analysis fMRI study. Neuroimage, 63(1), 262–271. http://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2012.06.045

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2013). Crossmodal and action-specific: neuroimaging the human mirror neuron system. Trends in Cognitive Sciences, 17(7), 311–8. http://doi.org/10.1016/j.tics.2013.04.012

Oosterhof, N. N., Wiggett, A. J., Diedrichsen, J., Tipper, S., & Downing, P. E. (2010). Surface-based information mapping reveals crossmodal vision–action representations in human parietal and occipitotemporal cortex. Journal of Neurophysiology, 104(2), 1077–1089. http://doi.org/10.1152/jn.00326.2010.

Oosterwijk, S., Winkielman, P., Pecher, D., Zeelenberg, R., Rotteveel, M., & Fischer, A. H. (2012). Mental states inside out: switching costs for emotional and nonemotional sentences that differ in internal and external focus. Memory & Cognition, 40(1), 93–100. http://doi.org/10.3758/s13421-011-0134-8

Oram, M. W., & Perrett, D. I. (1994). Responses of Anterior Superior Temporal Polysensory (STPa) Neurons to “Biological Motion” Stimuli. Journal of Cognitive Neuroscience, 6(2), 99–116. http://doi.org/10.1162/jocn.1994.6.2.99

Orban, G. A., & Caruana, F. (2014). The neural basis of human tool use. Frontiers in Psychology, 5(APR), 1–12. http://doi.org/10.3389/fpsyg.2014.00310

Orlov, T., Makin, T. R., & Zohary, E. (2010). Topographic representation of the human body in the occipitotemporal cortex. Neuron, 68(3), 586–600. http://doi.org/10.1016/j.neuron.2010.09.032

Orlov, T., Porat, Y., Makin, T. R., & Zohary, E. (2014). Hands in motion: an upper-limb-selective area in the occipitotemporal cortex shows sensitivity to viewed hand kinematics. The Journal of Neuroscience : The Official Journal of the Society for

Neuroscience, 34(14), 4882–95. http://doi.org/10.1523/JNEUROSCI.3352-13.2014

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. Cognition, 108(1), 155–84. http://doi.org/10.1016/j.cognition.2008.02.007

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The Two-Body Inversion Effect. Psychological Science, 1–11. http://doi.org/10.1177/0956797616685769

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. Psychological Science, 14(2), 119–124. http://doi.org/10.1111/1467-9280.t01-1-01429

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), 2096–107. http://doi.org/10.1162/jocn_a_00257

Peelen, M. V, Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. Neuron, 49(6), 815–822. http://doi.org/10.1016/j.neuron.2006.02.004

Peeters, R. R., Rizzolatti, G., & Orban, G. A. (2013). Functional properties of the left parietal tool use region. NeuroImage, 78, 83–93. http://doi.org/10.1016/j.neuroimage.2013.04.023

Peeters, R., Simone, L., Nelissen, K., Fabbri-Destro, M., Vanduffel, W., Rizzolatti, G., & Orban, G. A. (2009). The representation of tool use in humans and monkeys: Common and uniquely human features. Journal of Neuroscience, 29(37), 11523–11539. http://doi.org/10.1523/JNEUROSCI.2040-09.2009

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10(4), 437–442. http://doi.org/10.1163/156856897X00366

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., … Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. Nature Communications, 9(1). http://doi.org/10.1038/s41467-018-03068-4

Peuskens, H., Vanrie, J., Verfaillie, K., & Orban, G. A. (2005). Specificity of regions processing biological motion. The European Journal of Neuroscience, 21(10), 2864–75. http://doi.org/10.1111/j.1460-9568.2005.04106.x

Pinker, S. (1989). Learnability and Cognition: The Acquisition of Argument Structure. In Language (Vol. 68, p. xiv, 411 p.).

Potter, M. C. (1976). Short-Term Conceptual Memory for Pictures. Journal of Experimental Psychology : Human Learning and Memory, 2(5), 509–522.

Rissman, L., Rawlins, K., & Landau, B. (2015). Using instruments to understand argument structure: Evidence for gradient representation. Cognition, 142, 266–290. http://doi.org/10.1016/j.cognition.2015.05.015

Rizzolatti, G., & Craighero, L. (2004). the Mirror-Neuron System. Annual Review of Neuroscience, 27(1), 169–192. http://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature Reviews. Neuroscience, 11(4), 264–74. http://doi.org/10.1038/nrn2805

Roberts, K. L., & Humphreys, G. W. (2010). Action relationships concatenate representations of separate objects in the ventral visual system. NeuroImage, 52(4), 1541–8. http://doi.org/10.1016/j.neuroimage.2010.05.044

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. Current Biology, 23(3), 250–254. http://doi.org/10.1016/j.cub.2012.12.017

Romagno, D., Rota, G., Ricciardi, E., & Pietrini, P. (2012). Where the brain appreciates the final state of an event: the neural correlates of telicity. Brain and Language, 123(1), 68–74. http://doi.org/10.1016/j.bandl.2012.06.003

Ross, H. (1972). Play It Again, Sam. United States: Paramount Pictures.

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. Nature Neuroscience, 9(11), 1421–1431. http://doi.org/10.1038/nn1786

Salzman, C. D., Britten, K. H., & Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. Nature, 346(6280), 174–177. http://doi.org/10.1038/346174a0

Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. Brain : A Journal of Neurology, 130(Pt 9), 2452–61. http://doi.org/10.1093/brain/awm162

Scholl, B. J. (2001). Objects and attention: The state of the art. Cognition, 80(1–2), 1–46. http://doi.org/10.1016/S0010-0277(00)00152-9

Scholl, B. J., & Gao, T. (2013). Percieving animacy and intentionality: Visual processing or higher-level judgment? In M. D. Rutherford & V. A. Kuhlmeier (Eds.), Social perception: Detection and interpretation of animacy, agency, and intention. MIT Press.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, Development and "Theory of Mind." Mind and Language, 14(1), 131–153. http://doi.org/10.1111/1468-0017.00106

Schuler, K. K. (2005). VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Dissertation Abstracts International, B: Sciences and Engineering.

Schwarzlose, R. F., Swisher, J. D., Dang, S., & Kanwisher, N. (2008). The distribution of category and location information across object-selective regions in human visual cortex. Proceedings of the National Academy of Sciences of the United States of America, 105(11), 4447–52. http://doi.org/10.1073/pnas.0800431105

Scott, T. L., Gallée, J., & Fedorenko, E. (2016). A new fun and robust version of an fMRI localizer for the frontotemporal language system. Cognitive Neuroscience, 8928(July), 1–10. http://doi.org/10.1080/17588928.2016.1201466

Senior, C., Barnes, J., Giampietro, V., Simmons, A., Bullmore, E. T., Brammer, M., & David, A. S. (2000). The functional neuroanatomy of implicit-motion perception or representational momentum. Current Biology : CB, 10(1), 16–22.

Shiffrar, M., & Freyd, J. (1993). Timing and apparent motion path choice with human body photographs. Psychological Science.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review, 84(2), 127–190. http://doi.org/10.1037/0033-295X.84.2.127

Shirai, N., & Imura, T. (2016). Emergence of the ability to perceive dynamic events from still pictures in human infants. Scientific Reports, 6(November), 37206. http://doi.org/10.1038/srep37206

Simoncelli, E. P., Heegert, D. J., & Heeger, D. J. (1998). A Model of Neuronal Responses in Visual Area MT. Vision Research, 38(5), 743–761. http://doi.org/S0042698997001831 [pii]

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. ArXiv Preprint ArXiv:1406.2199, 1–11. http://doi.org/10.1017/CBO9781107415324.004

Singer, J. M., & Sheinberg, D. L. (2010). Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 30(8), 3133–45. http://doi.org/10.1523/JNEUROSCI.3211-09.2010

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage, 44(1), 83–98. http://doi.org/10.1016/j.neuroimage.2008.03.061

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. Developmental Science, 10(1), 89–96. http://doi.org/10.1111/j.1467-7687.2007.00569.x

Spence, C., Nicholls, M. E., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. Perception & Psychophysics, 63(2), 330–336. http://doi.org/10.3758/BF03194473

Strickland, B. (2016). Language reflects “core” cognition: A new theory about the origin of cross-linguistic regularities. Cognitive Science, 1–32. http://doi.org/10.1111/cogs.12332

Strickland, B., & Keil, F. (2011). Event completion: event based inferences distort memory in a matter of seconds. Cognition, 121(3), 409–15.

http://doi.org/10.1016/j.cognition.2011.04.007

Strickland, B., & Scholl, B. J. (2015). Visual Perception Involves Event-Type Representations: The Case of Containment Versus Occlusion. Journal of Experimental Psychology: General, 144(3), 570–580.

Talmy, L. (2000). Toward a Cognitive Semantics. Cambridge, MA: MIT Press.

Tarhan, L. Y., Watson, C. E., & Buxbaum, L. J. (2015). Shared and distinct neuroanatomic regions critical for tool-related action production and recognition: Evidence from 131 left-hemisphere stroke patients. Journal of Cognitive Neuroscience.

Taylor, J. C., Wiggett, A. J., & Downing, P. E. (2007). Functional MRI analysis of body and body part representations in the extrastriate and fusiform body areas. Journal of Neurophysiology, 98(3), 1626–33. http://doi.org/10.1152/jn.00012.2007

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. Nature. http://doi.org/10.1038/381520a0

Tikhonov, A. N., & Arsenin, V. Y. (1977). Solutions of Ill-Posed Problems. Washington, D.C.: V.H. Winston & Sons.

Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74(3), 209–253. http://doi.org/10.1016/S0010-0277(99)00069-4

Tootell, R. B., Reppas, J. B., Kwong, K. K., Malach, R., Born, R. T., Brady, T. J., … Belliveau, J. W. (1995). Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 15(4), 3215–30.

Trueswell, J. C. J. C., Medina, T. N. T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: fast mapping meets cross-situational word learning. Cognitive Psychology, 66(1), 126–56. http://doi.org/10.1016/j.cogpsych.2012.10.001

Trueswell, J. C., & Kim, A. E. (1998). How to Prune a Garden Path by Nipping It in the Bud: Fast Priming of Verb Argument Structure. Journal of Memory and Language, 39(1), 102–123. http://doi.org/10.1006/jmla.1998.2565

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. Journal of Memory and Language, 63(1), 64–82. http://doi.org/10.1016/j.jml.2010.02.006

Tucciarelli, R., Turella, L., Oosterhof, N. N., Weisz, N., & Lingnau, A. (2015). MEG multivariate analysis reveals early abstract action representations in the lateral occipitotemporal cortex. Journal of Neuroscience, 35(49), 16034–16045. http://doi.org/10.1523/JNEUROSCI.1422-15.2015

Urgesi, C., Candidi, M., & Avenanti, A. (2014). Neuroanatomical substrates of action perception and understanding: an anatomic likelihood estimation meta-analysis of lesion-symptom mapping studies in brain injured patients. Frontiers in Human Neuroscience, 8(May), 344. http://doi.org/10.3389/fnhum.2014.00344

Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. Proceedings of the National Academy of Sciences of the United States of America, 98(20), 11656–61. http://doi.org/10.1073/pnas.191374198

van Boxtel, J. J. A., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. Journal of Vision, 13(12), 1–16. http://doi.org/10.1167/13.12.7.doi

van Buren, B., Uddenberg, S., & Scholl, B. J. (2015). The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. Psychonomic Bulletin & Review. http://doi.org/10.3758/s13423-015-0966-5

Vangeneugden, J., De Mazière, P. A., Van Hulle, M. M., Jaeggli, T., Van Gool, L., & Vogels, R. (2011). Distinct mechanisms for coding of visual actions in macaque temporal cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 31(2), 385–401. http://doi.org/10.1523/JNEUROSCI.2703-10.2011

Vangeneugden, J., Peelen, M. V, Tadin, D., & Battelli, L. (2014). Distinct neural mechanisms for body form and body motion discriminations. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 34(2), 574–85.

http://doi.org/10.1523/JNEUROSCI.4032-13.2014

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. Journal of Cognitive Neuroscience, 13(4), 454–61.

Verfaillie, K., & Daems, A. (1996). The priority of the agent in visual event perception: On the cognitive basis of grammatical agent-patient asymmetries. Cognitive Linguistics, 7(1996), 131–148. http://doi.org/10.1515/cogl.1996.7.2.131

Watson, C., Cardillo, E., Ianni, G., & Chatterjee, A. (2013). Action concepts in the brain: an activation likelihood estimation meta-analysis. Journal of Cognitive Neuroscience, 25(8), 1191–1205.

Watson, C. E., Cardillo, E. R., Bromberger, B., & Chatterjee, A. (2014). The specificity of action knowledge in sensory and motor systems. Frontiers in Psychology, 5(May), 1–11. http://doi.org/10.3389/fpsyg.2014.00494

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses. PLoS ONE, 9(11), 1–19. http://doi.org/10.1371/journal.pone.0112575

Weiner, K. S., & Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. Psychological Research, 77(1), 74–97. http://doi.org/10.1007/s00426-011-0392-x

White, A. S., Reisinger, D., Rudinger, R., Rawlins, K., & Durme, B. Van. (2017). Computational linking theory.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. Behavior Research Methods, 42(3), 671–684. http://doi.org/10.3758/BRM.42.3.671

Williams, A. (2015). Arguments in Syntax and Semantics. Cambridge University Press. http://doi.org/10.1017/CBO9781139042864

Wilson, F., Papafragou, A., Bunger, A., & Trueswell, J. (2011). Rapid extraction of event

participants in caused motion events. In Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX.

Winawer, J., Huk, A. C., & Boroditsky, L. (2008). A Motion Aftereffect From Still Motion Photographs Depicting Motion. Psychological Science, 19(3), 276–83. http://doi.org/10.1111/j.1467-9280.2008.02080.x

Winawer, J., Huk, A. C., & Boroditsky, L. (2010). A motion aftereffect from visual imagery of motion. Cognition, 114(2), 276–284. http://doi.org/10.1016/j.cognition.2009.09.010

Windisch-Brown, S., Dligach, D., & Palmer, M. (2011). VerbNet class assignment as a WSD task. Proceedings of the 9th International Conference on Computational Semantics, 2712345267(2007), 85–94. http://doi.org/10.1007/978-94-007-7284-7

Woodard, K., Gleitman, L. R., & Trueswell, J. C. (2016). Two- and three-year-olds track a single meaning during word learning: Evidence for propose-but-verify. Language Learning and Development, 12(3), 252–261. http://doi.org/10.1080/15475441.2016.1140581

Wurm, M., Caramazza, A., & Lingnau, A. (2017). Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. Journal of Neuroscience, 37(3), 562–575. http://doi.org/10.1523/JNEUROSCI.1717-16.2017

Wurm, M. F., Ariani, G., Greenlee, M. W., & Lingnau, A. (2015). Decoding concrete and abstract action representations during explicit and implicit conceptual processing. Cerebral Cortex, 1–12. http://doi.org/10.1093/cercor/bhv169

Wurm, M. F., & Caramazza, A. (2018). Representation of action concepts in left posterior temporal cortex that generalize across vision and language, 0–25.

Wurm, M. F., & Lingnau, A. (2015). Decoding actions at different levels of abstraction. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 35(20), 7727–7735. http://doi.org/10.1523/JNEUROSCI.0188-15.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United

States of America, 111(23), 8619–24. http://doi.org/10.1073/pnas.1403112111

Yeul Bae, G., & Flombaum, J. I. (2011). Amodal causal capture in the tunnel effect. Perception, 40(1), 74–90. http://doi.org/10.1068/p6836

Yin, J., & Csibra, G. (2015). Concept-Based Word Learning in Human Infants. Psychological Science, 26(8), 1316–24. http://doi.org/10.1177/0956797615588753

Yuan, S., & Fisher, C. (2009). "'Really? She Blicked the Baby?"' Psychological Science, 20(5), 619–626. http://doi.org/10.1111/j.1467-9280.2009.02341.x

Yuille, A. L., & Liu, C. (2018). Deep Nets: What have they ever done for Vision? ArXiv, 1–19.

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. Journal of Experimental Psychology. General, 138(2), 307–327. http://doi.org/10.1037/a0015305

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. Psychological Bulletin, 133(2), 273–93. http://doi.org/10.1037/0033-2909.133.2.273

Zheng, M., & Goldin-Meadow, S. (2002). Thought before language: How deaf and hearing children express motion events across cultures. Cognition, 85(2), 145–175. http://doi.org/10.1016/S0010-0277(02)00105-1