NeuroImage 157 (2017) 511–520

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Neural representations of the concepts in simple sentences: Concept activation prediction and context effects

Marcel Adam Just*[,1], Jing Wang[1], Vladimir L. Cherkassky

Center for Cognitive Brain Imaging, Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Keywords:
Neural representations of concepts
Predictive modeling
Multi-concept sentences
FMRI
Sentence context effects

ABSTRACT

Although it has been possible to identify individual concepts from a concept's brain activation pattern, there have been significant obstacles to identifying a proposition from its fMRI signature. Here we demonstrate the ability to decode individual prototype sentences from readers' brain activation patterns, by using theory-driven regions of interest and semantic properties. It is possible to predict the fMRI brain activation patterns evoked by propositions and words which are entirely new to the model with reliably above-chance rank accuracy. The two core components implemented in the model that reflect the theory were the choice of intermediate semantic features and the brain regions associated with the neurosemantic dimensions. This approach also predicts the neural representation of object nouns across participants, studies, and sentence contexts. Moreover, we find that the neural representation of an agent-verb-object proto-sentence is more accurately characterized by the neural signatures of its components as they occur in a similar context than by the neural signatures of these components as they occur in isolation.

Introduction

Concepts may be the basic building blocks of thought, but the minimally composed structure of human thought is a proposition consisting of multiple concepts. We report here the capability of predicting the brain activation patterns evoked by the reading of an agent-verb-object proto-sentence. We develop a model that estimates the activation pattern of component words from the mappings learned in context-sensitive environments, and combines them to produce predictions of the resulting activation.

The types of concepts that have previously been most amenable to a mapping between a stimulus item and a brain activation pattern have been concrete object concepts. This type of mapping was initially demonstrated in studies in which the objects were presented visually or were being recalled (Carlson et al., 2003; Connolly et al., 2012; Cox and Savoy, 2003; Eger et al., 2008; Hanson et al., 2004; Haxby et al., 2001; Ishai et al., 1999; Mitchell et al., 2004; O'Toole et al., 2005; Polyn et al., 2005; Shinkareva et al., 2008), and subsequently in studies where the concept was evoked by the word that named it (Just et al., 2010; Peelen and Caramazza, 2012; Shinkareva et al., 2011). Predictive modeling of brain activity associated with concepts was enabled by the postulation of a mediating layer of perceptual and semantic features of the objects, resulting in the decoding from their fMRI signature pictures or text concerning objects (Anderson et al., 2015; Mitchell et al., 2008; Pereira et al., 2011), natural images (Naselaris et al., 2009), faces (Cowen et al., 2014), objects and actions in video clips (Huth et al., 2012; Nishimoto et al., 2011) and in speech (Huth et al., 2016). Other studies have found distinct activation patterns associated with the neural representations of concepts of varying degrees of semantic abstractness (Anderson et al., 2014; Ghio et al., 2016; Wang et al., 2013). A few studies have further demonstrated the ability to associate brain activation patterns with inter-concept relations in a proposition (Frankland and Greene, 2015; Wang et al., 2016). However, characterizing the neural representations of sentences and the effect of contexts has remained a considerable challenge.

The major advance attempted in the current study was to characterize the neural representation of simplified prototype sentences and the effect of context within a theory-driven computational framework. The model was theory-driven, built on the previous knowledge of the neural representations of objects in several ways. First, the stimulus sentence prototypes (e.g., *Plumber grabs pliers*) described a scenario associated with a general theme pertaining to one of three known semantic dimensions of neural representation, namely *shelter, manipulation, or eating*. Second, the meaning components of the stimulus word-concepts were defined as the concepts' relatedness to each of the three dimensions. These three semantic properties were used to predict

* Corresponding author.
E-mail address: just@cmu.edu (M.A. Just).
[1] These authors made equal contributions to the work.

http://dx.doi.org/10.1016/j.neuroimage.2017.06.033
Received 21 November 2016; Accepted 15 June 2017
Available online 17 June 2017
1053-8119/

Table 1
Thirty-six stimulus sentences. Each set includes 9 sentences that are composed from 27 content words.

	Set 1	Set 2	Set 3	Set 4
Shelter	Explorer enters car	Explorer repairs church	Explorer exits house	Tourist enters car
	Hiker exits church	Hiker enters house	Hiker repairs car	Explorer exits church
	Tourist repairs house	Tourist exits car	Tourist enters church	Hiker repairs house
Manipulation	Plumber drops hammer	Plumber lifts knife	Plumber grabs pliers	Mechanic drops hammer
	Carpenter grabs knife	Carpenter drops pliers	Carpenter lifts hammer	Plumber grabs knife
	Mechanic lifts pliers	Mechanic grabs hammer	Mechanic drops knife	Carpenter lifts pliers
Eating	Diner bites carrot	Diner tastes celery	Diner chews tomato	Picnicker bites carrot
	Glutton chews celery	Glutton bites tomato	Glutton tastes carrot	Diner chews celery
	Picnicker tastes tomato	Picnicker chews carrot	Picnicker bites celery	Glutton tastes tomato

the neural representation of word concepts and the proto-sentences that they composed. Third, the analysis of the activation patterns focused on the *a priori* specified brain regions that were associated with these three dimensions. The approach was computational in the sense that it established the mapping between the concepts' semantic properties on the three dimensions and the voxel activation patterns associated with reading the sentences. This mapping was sufficiently detailed to characterize the behavior of individual voxels with respect to the neurosemantic properties, and sufficiently robust to be generalized to predict the neural signatures of new sentences with similar contexts but composed of new words.

A brief terminological and theoretical disclaimer is warranted about our use of the word *sentence* in this article. The stimuli were proto-sentences, in that they were lacking articles for the nouns: *Hiker enters house* was presented instead of *The hiker enters the house*. The stimuli would be more accurately referred to as *proto-sentences*, but for brevity following this disclaimer we use the term *sentences*. The theoretical disclaimer concerns the fact that our analysis and model focuses on the neural representation of the three content words of each sentence, and not their thematic or syntactic structure. In fact, all of the stimuli were identical with respect to thematic and syntactic structure (agent-verb-object), and our model does not characterize the representation of these identical aspects of the sentences. Thus our reference to the neural representation of the sentence stimuli refers to the neural representation of the three content words, and not to thematic or syntactic aspects of the representation.

We hypothesized that the neural activation pattern evoked by a simple proto-sentence can be predicted by the semantic properties of the sentence's component words. The theoretical background for construing the neural representations of concrete objects is a neural/semantic account of concrete object representation (Just et al., 2010; Mitchell et al., 2008) that postulates that the neural signature of a concept is composed of component activations in the various brain subsystems that come into play during the consideration of, or interaction with, the concept. These component activations are associated with different dimensions of the meaning of the concept. For example, thinking about a *knife* evokes motor and premotor areas, which have been associated with action verb processing in previous studies (Hauk et al., 2004; Pulvermüller et al., 2005). These regions are found to respond to various manipulable objects such as hand tools, and to not respond to non-manipulable objects, indicating their role in representing the meaning dimension of *manipulation* or body-object interaction. The three main neurosemantic dimensions underlying the representation of 60 concrete nouns were *shelter, manipulation,* and *eating* (Just et al., 2010). Such sets of dimensions and their corresponding brain subsystems are proposed to constitute part of the basis set for neurally representing concrete objects. Thus, the activation pattern associated with a particular concrete concept should be predictable, based on the concept's semantic properties that are encoded by these brain subsystems.

Of course, a sentence context is very likely to modulate the neural representation of its component concepts, but the neural modulatory principles by which such context effects operate have not been determined. We propose a hypothesis and a method to decode the multiple concepts embedded in a sentence context from a particular form of their fMRI activation signature. One key to the method is to base the estimate of a concept's neural signature on its instantiation in a set of roughly similar sentence contexts (excluding the sentence whose activation is being predicted). This approach assumes that the neural signature of a component concept in context is modulated by the context, and thus differs systematically from the signature of the word when it is processed in isolation, an assumption that we tested. The estimate of the neural signature of a concept was obtained by averaging the activation patterns of several different sentences containing the concept, on the assumption that the neural signals contributed by the other concepts in the sentences would be cancelled out by the averaging. The implication underlying this method is that the neural representation of a simple proto-sentence can be predicted by the sum of the neural representations of its content word concepts, as estimated from approximately similar contextual environments.

Materials and methods

Materials

Each of the 36 simplified three-word sentences (of the form *Agent-verb-object*) described a scenario associated with a general theme pertaining to one of three semantic factors *shelter, manipulation, or eating*, previously shown to underlie the representations of concrete nouns (Just et al., 2010). The objects were selected from among the 60 concrete objects whose associated activation patterns were factor analyzed in a previous study (Just et al., 2010); the selected objects were those with some of the highest factor scores for their dominant semantic factor. The agents and verbs were chosen for consistency with the theme or factor. There were 12 sentences per theme, for a total of 36 sentences, as shown in Table 1. The sentences were composed of triplets of words from among 27 content words: 9 words of each word class (agent, verb, and object). Within each class, 3 words were associated with each neurosemantic factor. The 36 sentences were assigned to four sets, so that each set contained nine sentences, three per theme, and sentences in each set used all 27 words, each occurring in one sentence. Each of the four blocks of trials presented the 36 sentences in a different order.

Participants

Ten healthy, right-handed, native speaking adults (7 females) from the Carnegie Mellon community participated in the fMRI experiment. All participants gave written informed consent approved by the Carnegie Mellon Institutional Review Board. The data from all the participants were included in the analyses below.