THE INFLUENCE OF SEMANTICS ON THE VISUAL PROCESSING OF NATURAL SCENES

BY

MANOJ KUMAR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Neuroscience
in the Graduate College of the
University of Illinois at Urbana-Champaign 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Diane M. Beck, Chair
Professor Kara D. Federmeier
Professor Gabriele Gratton
Assistant Professor Oluwasanmi Koyejo

Abstract

A long standing question in cognitive science has been: is visual processing completely encapsulated and separate from semantics or can visual processing be influenced by semantics? We address this question in two ways: 1) Do pictures and words share similar representations and 2) Does semantics modulate visual processing. Using multi-voxel pattern analysis (MVPA) and fMRI decoding we examined the similarity of neural activity across pictures and words that describe natural scenes. A whole brain MVPA searchlight revealed multiple brain regions in the occipitotemporal, posterior parietal and frontal cortices that showed transfer from pictures to words and from words to pictures. In addition to sharing similar representations across pictures and words, can words dynamically influence the processing of visual stimuli? Using Event Related Potentials (ERPs) and good and bad exemplars of natural scenes, we show that top-down expectation, initiated via a category cue (e.g. the word 'Beach'), dynamically influences the processing of natural scenes. Good and bad exemplars first evoked differential ERPs in the time-window 250-350 ms from stimulus onset, with the bad exemplars showing greater negativity over frontal electrode sites, when the cue matched the image. Interestingly, this good/bad effect disappeared when the images were mismatched to the cue. Overall, these studies taken together, provide evidence for the influence of semantics on the visual processing of natural scenes.

Acknowledgments

I have been quite fortunate in my journey into neuroscience and I have many people to be thankful for. Doing a PhD has been a fun, long, and arduous journey where the light at the end of one tunnel could be emanating from the entrance of a different tunnel, and sometimes one never escapes these tunnels. It is even more challenging when the journey involves a significant career change (or hara-kiri as some may call it) and relocating family. It's taken a village (or three), spanned across continents, to make all this happen and it is to many that I am indebted to. It is on the shoulders of these giants that I have stood and I would like to thank them all.

My parents, brother, bhabi and my extended family for tolerating my crazy pursuits (and often being skeptical) of doing research and going off the beaten path. My dad did some remarkable things to raise and support everyone in the family, with very limited resources. His deeds always inspire me. He taught to me to be fearless and be unfettered by obstacles. He truly lived the phrase -- Impossible is nothing. My mom always encouraged me to read, even though we had to pay to borrow books at the local library. She has been the one constant source of encouragement for me to attend college and pursue non-lucrative careers. I know she did this despite considerable hardships she has faced through her life.

Veena and Moksh are the love and strength of my life. Veena's family has been supportive (and skeptical at the same time) of my endeavors. I greatly cherish all the support to Veena, Moksh and me through all these years from everyone in her family. Moksh has been a delight to spend time with in the course of my PhD. Most importantly, I thank Veena. Her encouragement, support and

the sacrifices that she has made for me to pursue this path have been tremendous. Without her, this journey would have been impossible.

I got to study at some of the best schools in Bangalore: St. Germains, and Bishop Cottons. Triumph through Trials (Ad Augusta, Per Angusta) was the school motto (St. Germains) and so is true of life and science.My teachers in my high school period: Mrs. Aranha, Mr. Murthy, Ms. Alurkar, Ms. Maithili, Mr. Francis Matthew, Mr. Andrews and Mr. Chandran were wonderful teachers who inspired a love for learning. The National Geographic magazine, in the school library, played a great role in fostering a love for science. It was a rare commodity in those days and it opened my view into a world of advanced science and technology. In addition to these teachers, my many friends at school have been invaluable, I had many wonderful times spent discussing all the beauty of math and physics during these years with them.

I have had many things to cherish during my undergrad years at IT-BHU. The time spent in all the extracurricular activities there have helped me in more ways than I imagined in both my careers in science and industry. The friends I made there continue to be some of the most supportive through all these years.

I have been fortunate to work in some labs that foster a great research environment. It was with Dr. P.N. Shankar at the National Aerospace Laboratories, Bangalore, that I got my first exposure to a world class fluid dynamicist and applied mathematician. He has been a greater mentor throughout my career and an inspiration with his outstanding contributions to fluid dynamics and astronomy. By working with him I learnt a lot about what it takes to do great science, especially in India. His

interests span many fields apart from his specialty in fluid dynamics. It was on a visit to his house, fifteen years after I worked with him, that I picked up a book from his bookshelf, called Phantoms in the Brain, and decided to quit my career in industry and pursue research in cognitive neuroscience.

My friends Nagarajan Sankrithi and Siddharth Panda for being great sounding boards during my early days in the US and for some great cooking lessons.

From my days at Yale, I met many wonderful people. Hitten Zaveri at Yale, for being a great friend, mentor and encouraging me to pursue neuroscience. K. R. Sreenivasan for being a great mentor and helping me navigate many waters at Yale. He even helped me when he was at Trieste and now NYU. Juan Fernando de la Mora for one of the best math courses I have ever taken. My many friends at Yale Engineering, Computer Science and MB&B who supported me through many ups and downs at Yale. A big thanks to all the people who I worked with at ASHA-Yale. These friendships have continued to play a big role in my life. I have many fond memories of time spent with Stefano, Bratin, Rajdeep, Kakoli, Iban, Brindesh, Chris, Sudha and Kostas.

Mike King at the University of Michigan for taking me on in his lab and mentoring me into neuroscience. This was in a period when I could only work on weekends or late evenings and Mike was willing to work with me on this schedule.

The University of Illinois at Urbana-Champaign is an outstanding school and the Neuroscience Program helped smoothen my transition into the PhD program. Beginning at the interview stage to

enter the program to finishing my PhD, I have had great support and mentoring. It's been an absolute pleasure to work with and learn from so many wonderful people here: Sam Beshers, Bob Wickesberg, Doug Jones, and Neal Cohen played a big role during my interviews and when I started my graduate work. Also, a big thanks to my colleagues in the IGERT program. They all helped make my transition easier.

My advisor Diane Beck and co-advisor Kara Federmeier who have both been great pillars of support and encouragement. They have helped me enormously in the trials and triumphs of the PhD and have helped me grow as a scientist. They have created a fantastic research environment in their labs from which I have benefitted. I have learnt a lot from them and owe them a great debt of gratitude. I hope to emulate the high standards of scientific and human engagement that they set.

In addition to my advisors, other faculty have been instrumental in my success. Daniel Simons, Gabrielle Gratton, Derek Hoeim and Oluwasanmi Koyejo have been on my committees. I have also taken some great courses from Dan Roth, Alejandro Lleras, and Aaron Benjamin.

My fellow grad students and lab mates have provided great support and a rich intellectual environment through all the PhD years. In Diane's lab: Eamon, Audrey, Brian, Evan, Ramisha, and Heeyoung. In Kara's lab: Mallory, Danielle, Michelle, Cybelle, Joost, Brennan, Katie and Dan. In addition, the Viscog lab meetings have been a great forum to discuss science, research methods, and non-science matters. I have also gotten great feedback and advice on my conference presentations and job talks from this group.

Fei-Fei Li at Stanford has been a good sounding board and a source of encouragement on taking problems at the boundary of semantics and vision. Chris, Catalin, and Michelle have been wonderful collaborators providing advice, feedback, and materials on many projects.

I have been inspired by the work and writings of many scientists that I have never met but read have about. The greatest inspirations have come from Raman, Feynman, and Hilbert.

I would also like to acknowledge the many lessons that failure has taught me. I have failed multiple times at various points in my career. Failures are always hard to deal with in the moment but looking back, all the dots of failure have miraculously connected to build a path towards my PhD.

This work has been supported by several funding agencies. I would like to thank them for their generous support: National Science Foundation IGERT Fellowship (MK), James S. McDonnell Foundation Scholar award (KDF), National Institutes of Health Grant 1 R01 EY019429 (LFF and DMB) and ONR MURI (LFF and DMB).

Table of Contents

Chapter 1: Introduction…….....1

Chapter 2: Evidence For Similar Patterns of Neural Activity Elicited by Picture- and Word-based Representations of Natural Scenes…….....11

Chapter 3: The Good Bad Effect: How Does Representativeness Inform Vision?.....38

Chapter 4: Expectancy Modulates the Good-Bad Effect……..…57

Chapter 5: Conclusion…….....76

References……..81

Appendix A: List of Phrases Describing Natural Scenes……...…...…….....91

Appendix B: Cross-Decoding Map: 400 Voxel Cluster……...…….……...100

Appendix C: Confusion Matrices and BOLD measurements for ROIs……...101

Appendix D: Cross-Decoding Maps In Each Direction…….…..107

Chapter 1

Introduction

When you view a picture of a beach, you evoke some concept of a beach in your mind. You also evoke a concept of a beach when you read the phrase "Beautiful Beach". These concepts evoked by viewing pictures or reading words, share some similarities, even though they are elicited by completely different inputs —one is a pictorial input and the other a verbal input. Moreover, these input modalities not only evoke similar concepts, but can also dynamically influence each other. If I gave you the verbal cue of "umbrella" prior to showing you a picture of an umbrella, you would expect to see an umbrella held open vertically (as it is typically used). It is unlikely you would expect to be shown an umbrella held open horizontally or to see someone standing under an umbrella while deep inside a swimming pool. You also might expect to see an umbrella that is black in color, and not aquamarine. You have thus used the semantics of an umbrella, accessed via the verbal cue, to set up expectations of what you are likely to see -- for example it's shape, size, color, orientation and probable location. Do these expectations, set up via semantics, influence visual perception? If they do influence visual perception, how do they do so? These are the central questions of this dissertation.

The question of whether, when, and how semantics affects vision is a long-standing one. One definition of the semantics of an item, is all knowledge that constitutes our concept of that item (Cree & McRae, 2003; McRae, Cree, Seidenberg, & Mcnorgan, 2005). It encompasses information about the item across multiple modalities (visual, auditory, touch, taste and smell) and can include abstract information about the item e.g. "beautiful" or even encyclopedic information ("zebras are found in Africa"). It is well established that visual processing can lead to semantic processing; for example, after viewing a picture of a beach, we evoke a concept of a beach in our minds. What is under debate is whether semantics or prior knowledge can feedback and modulate visual processing and perception (Pylyshyn, 1999; Firestone and Scholl, 2014). In this thesis, I am going to address this question in two ways. First, I will use functional magnetic resonance imaging (fMRI) along with multivoxel pattern analysis (MVPA) (Norman, Polyn, Detre, & Haxby, 2006) techniques to examine whether there are similar patterns of brain activation for semantic category information that is accessed through pictures and through verbal stimuli. If we do find evidence for a similar representation across pictures and verbal stimuli in

local brain regions, and we already have some evidence (discussed below in Section 1.1) albeit not at a fine local scale, this would imply that the neural code is evoked similarly, in localized regions, for pictures and verbal stimuli and that semantic access is independent of the modality of the stimulus in these regions. Thus, these regions could serve as potential sites where we may see interaction in processing across modalities. So my second method to determine the influence of semantics on visual processing is: given that semantic information, across modalities (pictures and words), accesses the same network of brain regions, can semantic information, activated via words, impact pictorial processing. I will use time sensitive measures – event-related potentials (ERPs) to test whether semantics, accessed via a verbal cue, can quickly modulate aspects of visual processing, and, if so, hone in on when in the course of processing such effects arise.

Our understanding, thus far, of how semantic information is stored and processed in the brain, across modalities, has been developed from a wide variety of studies spanning lesion studies, behavioral measures, electrophysiological measurements and neuroimaging. To gain a better understanding of the underlying semantic representations in the brain, I review the literature on semantics across modalities in **Section 1.1**. The neural mechanisms of feedback processing in the visual system, that can serve as pathways for semantics to interact with visual processing are discussed in **Section 1.2**.

1.1. Processing pictures and words in the brain.

We do evoke similar concepts when viewing a picture (eg. picture of a beach), or reading the word (eg. beach). It is natural to ask if these two different input types, pictures and written words (or even words listened to), are processed similarly to give rise to a common concept. Research on processing these input types has led to two dominant views in the literature: the multi-code view (Paivio, 1974; Warrington and Shalice, 1984), wherein pictures and words are processed in completely different ways with different memory stores and different representational codes; and the common code view, which suggests that pictures and words have common memory stores and have shared representational codes (Pylyshyn, 1973; Caramazza, Hillis, Rapp, & Romani, 1990). Recent work (discussed below), from measures of brain activity, is converging on the

common code view. This provides a framework to examine the interaction of semantics on visual processing, as the multi-code processing mechanism, wherein processing meaning from words and pictures constitute two completely separate systems, would allow minimal interactions between the two processing systems (Paivio, 1974). Thus, the possibility of a common code framework, indicating similar patterns of activity across pictures and words, will also allow for interaction between these two systems. The current research points to similar patterns of neural activity between pictures and words and we now examine the evidence, from Event Related Potentials (ERPs) and functional Magnetic Resonance Imaging (fMRI), that supports this viewpoint.

ERPs provide a continuous and instantaneous measure of electrical activity in the brain (Münte, Urbach, Düzel, & Kutas, 2000) and have helped us better understand cognitive processing in the brain. The N400 ERP component (Kutas and Hillyard, 1980), is known to index semantic access across a variety of modalities (words, sentences, pictures, sounds, gestures, cartoons), with stimuli congruent ("*sugar*") with the semantic context being facilitated (*"I take my coffee with cream and sugar"*) and stimuli that are incongruent (*"dog"*) to the semantic context are not facilitated (*"I take my coffee with cream and dog"*), resulting in greater negativities, in the N400 amplitude, for incongruent stimuli in (see review Kutas and Federmeier, 2011). The time-window when the semantics or meaning is extracted from a stimulus is processed has been determined to be in the time-window of 350-500 ms (Kutas and Hillyard, 1980; also see review Kutas and Federmeier, 2011). The N400 is thus well suited to study the time-course of semantic processing in different modalities and determine if similar patterns of responses occur when semantics is accessed by different stimulus modalities or input types (e.g. pictures and words). If a common-code exists, the pattern of N400 should have similarities across modalities or input types. Indeed, a variety of studies show a similarity in the pattern of the N400 amplitude: for sentences (Kutas and Hillyard, 1980); written words (Kutas, 1993); for written sentences with the last words sometimes replaced with a picture (Federmeier and Kutas, 2001; Nigam et al., 1992); spoken sentences with pictures (congruent or incongruent) simultaneously displayed when the critical noun (congruent or incongruent) is manipulated (Willems, Özyürek, & Hagoort, 2008); static images (Holcomb & Mcpherson, 1994); and video clips (Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008). Given the same pattern of N400 amplitude responses

to semantic processing under these different modalities, ERPs provide strong evidence that a common code exists across modalities. There are some scalp distributional differences between modalities, for example the N400 for words is distributed centro-parietally (Kutas, 1993), while for pictures it is fronto-central (Holcomb & Mcpherson, 1994; Federmeier and Kutas, 2001). Also, for pictures studies often show an earlier component, the N300, under certain conditions, that is related more to the form of the stimulus (see for example Holcomb & Mcpherson, 1994; Schendan and Kutas, 2002). Nevertheless, the common pattern of the N400 amplitude and it's sensitivity to meaning, does provide evidence for a common code across modalities and input type.

Neuroimaging, with its capability to localize brain regions, has provided some insights into the common code. A few studies have used paradigms with both picture and word stimuli, in the same experiment, to find brain areas that respond in common to pictures and words when these are processed for semantics (Kherif, Josse, & Price, 2010; Price, 2000; Vandenberghe, Price, Wise, Josephs, & Frackowiak, 1996). Regions of the left inferior frontal gyrus, the left fusiform gyrus and the angular gyrus have been shown to co-activate for pictures and words. In an experiment in which participants either viewed pictures of tools and animals or read words or text about tools and animals (Chao et al., 1999), the brain regions not only overlapped across pictures and words, but they also showed a more fine-grained similarity in terms of which categories activated which part of the brain. The ventral fusiform region showed activity for animals (pictures or words) and the posterior lateral temporal region showed activity for tools (pictures or words). Thus, these studies provide evidence that some brain regions show a common activation across pictures and words.

Common activity in a brain region, however, does not imply common representation, as the underlying neurons in these areas could process information differently for the different modalities. A better technique to find common code is to examine the patterns of neural activity for the two modalities. Two approaches have been used to perform MVPA (Norman, Polyn, Detre, & Haxby, 2006): cross-decoding (Kaplan et al., 2015) and representational similarity analysis (RSA) (Nili et al., 2014). In cross-decoding, a classifier is trained on the BOLD signal from one input type (eg. pictures) and then predicts the category of the stimulus from the BOLD

signal in another input type (eg. words). If cross-decoding is successful, the representations across the input types share a common hyperplane separating categories, and thus are similar. We can consider this to be a first order isomorphism (Kriegeskorte, Mur, & Bandettini, 2008) between the different input types. An alternative technique, RSA, uses the BOLD signal and computes distances between categories by computing a distance metric between signals. This distance is then compared to a similar distance metric computed from the signal in the other input type. If the metrics computed across input types are highly correlated, then the underlying neural representations across input type must be similar. This comparison provides a second order isomorphism between input types as not only are distances for a category across input types are computed, but also the distances between categories across input types. These two MVPA techniques thus can provide a more accurate inference on the existence of a common code as compared to univariate analysis.

Using cross-decoding, studies have shown common patterns of activity in the categorization of pictures of objects and nouns describing them (Fairhall and Caramazza, 2013; Shinkareva et al., 2011). This common pattern of activity is seen in frontal regions, regions associated with higher order visual processing and the angular gyrus. Using RSA (Devereux et al., 2003; Liuzzi et al., 2015; Bruffaerts et al., 2013) on object stimuli, representational similarity has been found in multiple brain regions across a variety of input types (written words, auditory words and pictures) thus providing a second order isomorphism between pictures and other input types for object categories. As compared to univariate analysis, these results provide stronger evidence for the existence of a common code for pictures and words in some brain regions.

Despite the use of MVPA techniques, there are some limitations in the existing literature in showing similar patterns of activity across input types (eg. pictures and written words). The stimuli chosen, for example animals versus tools, have preferential activity in separate brain regions (see Chao et al., 1999). Thus, decoding these categories in a region could be accomplished by a classifier even if the underlying neural code was dissimilar across input types, the mean signal for each category would be sufficient for the decoder to perform the task. What is lacking is showing similarity at a high local resolution, where different categories co-activate a local region. In this case, the existence of similar patterns of activation across input types must

necessarily be resolved in a higher dimensional space and thus provide a stronger measure of a common code across input types. Furthermore, little is known about the semantics of natural scenes as most studies have focused on objects (Devereux et al., 2013; Liuzzi et al., 2015; Bruffaerts et al., 2013; Fairhall and Caramazza, 2013; Shinkareva et al., 2011). Natural scenes are not only processed differently than objects (Oliva and Torralba, 2001; Greene and Oliva, 2009), they also co-activate local regions such as the parahippocampal place area (PPA) and retrosplenial cortex (RSC) across categories (Epstein and Kanwisher, 1998). Thus, they can prove to be good a stimulus set to check for the existence of a common code at better spatial resolution as compared to object categories.

In addition to knowing that a common-code across input types exists in a brain region, a related question is what is the nature of representation in regions where we do find a common code (Martin, 2007; Patterson et al., 2007). Understanding the underlying representation will give us better insights into which regions can contribute to the interaction of semantics and visual processing ongoing. One view is that semantic information is represented in a completely amodal format, abstracted from the input type (Lambon Ralph & Patterson, 2008; Patterson et al., 2007). Another view is that semantic information is stored as a combination of sensory and functional attributes in sensory motor systems that preferentially represent each sensory feature (Chao et al., 1999; Martin, 2007; Pulvermüller & Fadiga, 2010). The MVPA studies have found common code regions in sensory motor systems, as well as in areas not traditionally attributed to sensory motor systems, such as the angular gyrus (Devereux et al., 2013; Liuzzi et al., 2015; Bruffaerts et al., 2013; Fairhall and Caramazza, 2013; Shinkareva et al., 2011). Thus, the data suggests that common-code regions can span sensory-motor regions as well as regions where information can be abstracted away from the stimulus input type. The question does remain, what is the nature of the representation in sensory-motor systems, that do represent a common-code, is it abstract or sensory-motor specific?

In this section, I have reviewed the evidence for the existence of a common-code for pictures and words. MVPA studies (Devereux et al., 2013; Liuzzi et al., 2015; Bruffaerts et al., 2013; Fairhall and Caramazza, 2013; Shinkareva et al., 2011) have gone beyond showing just overlapping brain regions and have provide a better measure of the common-code. These studies have used simple

objects as their stimuli. In my thesis, I extend this work using natural scenes (Chapter 2). Also, these studies show multiple brain regions that represent information as common codes. These regions could potentially serve as sites for the interaction of semantics with visual processing. We next review the possible dynamics of the interaction of semantics with visual processing.

1.2 The Dynamics of Semantic Processing

To better understand the dynamics of the influence of semantics on visual processing, in this section we review the current state of what we know about the time-scale of semantic processing and the possible neural pathways that could enable its interaction with visual processing. Semantic processing in the brain is dynamic and is modulated by the context we are in, for example in the frame of reference of humans being attacked, we may put a snake and a bear in the same group, but if we are making size judgements, we will put the snake and bear in separate groups. Thus our context influences what features of the stimulus we make relevant under a particular context, and this choice can be influenced dynamically. The question remains, does this semantic information interact with visual processing? If so, at what time scale? These aspects of semantic processing are less understood and we examine what we currently know about the timescales of semantic processing and the dynamic neural pathways that could enable semantics to interact with visual processing.

At what timescales can semantics and visual processing interact? Using ERPs, particularly the N400 component, we have reliable and precise estimates of the time-course of semantic processing. The time-window when the semantics or meaning is extracted from a stimulus (words, sentences, pictures, sounds, gestures, cartoons) is processed has been determined to be no later than the time-window of 350-500 ms (Kutas and Hillyard, 1980; also see review Kutas and Federmeier, 2011). The time-scales at which visual processing takes place is approximately 0-300 ms from stimulus onset (Luck and Kappenman, 2011), starting with visual evoked potentials, that are sensitive to low level visual properties such as contrast, luminance and spatial frequency, in the following time-windows: the C1 50-70 ms, the P1 ~100 ms and the N1 ~130 ms (Schechter et al., 2005). Attentional mechanisms can modulate some of these visual processes, with the P1 and N1 components showing modulation in the time-window 100-250 ms

(Gonzalez, Clark, Fan, Luck, & Hillyard, 1994). Visual object processing is reflected in the N170 component for faces (Carmel and Bentin, 2002) and in the N250 component that indexes the identification of familiar objects (Scott, Tanaka, Sheinberg, & Curran, 2006). The larger scale structure of pictorial stimuli is indexed by the N300 component in the time-window 200-350 ms (Schendan and Kutas, 2002). Given these time-windows, 0-350 ms for visual processing and 350-500 ms for semantic processing, there are a few time points at which semantic processing and visual processing could interact. The first possibility is that semantic and visual processing occur in parallel and their interactions would occur during the course of their mutual processing; this type of interaction could occur anywhere in the time-window from 100-500 ms. The second possibility is that semantic cues, based on the context of the experimental situation, could pre-activate expected visual properties and modulate the processing of the incoming stimulus. It is this latter case that is of special interest as semantic evaluation of the incoming stimulus has not started, but expected semantic features, spanning modalities and input types, have been pre-activated by semantic context, initiated by words that do not bear any perceptual resemblance to the pictures, and the question remains can these pre-activated features modulate visual processing of the incoming stimulus, across the input type. This modulation of visual processing would need to occur in the first 350 ms, as visual processing has not been completed by then. There is already some evidence that this early modulation of visual processing occurs: semantic information can set up an upstream expectancy for perceptual features as early as 200 ms into processing (Federmeier and Kutas, 2001). In this thesis, I will use this paradigm (with no precuing in Chapter 3 and with precuing a scene category in Chapter 4) to answer the question: Can semantics modulate visual processing?

For semantics to modulate visual processing, there must exist pathways from semantic processing regions to visual regions or mechanisms for visual regions to be pre-activated by semantics. We discussed the existence of semantic regions, that process information independent of modality (Section 1.1). Current neuroimaging studies implicate several brain regions in the processing of semantic information. The angular gyrus, regions of the left lateral and ventral temporal cortex anterior to visual associative regions, left dorso-medial prefrontal cortex, left inferior frontal gyrus, left ventro-medial prefrontal cortex, and the posterior cingulate gyrus have been found to process semantic information for verbal stimuli ((Fairhall and Caramazza, 2013;

Binder et al., 2009). How are these semantic regions connected to modality specific regions? We do not yet have all the details of how information flows between a subset of these regions, or the nature of any recurrent interactions within each region. In the case of pre-activating semantics via a verbal cue, the cue can pre-activate expected features higher visual areas as these regions have shown to be multi-modal in nature. Thus, feedback between semantics and visual processing can occur at localized regions in the visual processing stream. An alternate mechanism for semantic preactivation to modulate visual processing, is through the tuning of a population of neurons to a particular category, for example if we attend to animals then multiple brain regions tune themselves to select features from animal categories as opposed to anything else (Çukur, Nishimoto, Huth, & Gallant, 2013). This mechanism ensures that multiple brain regions are ready to process information that has been pre-activated and could also cause extra processing to occur, if a stimulus other than the expected is shown. We study these effects in Chapter 4, where we use a paradigm wherein we match or mismatch the stimulus to the precue and observe the differences in the ERP waveforms to these conditions.

In addition to feedback pathways for pre-activated semantics, there are some plausible pathways for real time information transfer between semantic regions and visual areas. We do know that some semantic selection happens in the pre-frontal cortex (Martin, 2007; Binder, Desai, Graves, & Conant, 2009). The pre-frontal cortex could receive information very rapidly after stimulus onset via magnocellular pathways (Kveraga et al., 2007) and provide feedback through direct and cascading connections from the pre-frontal cortex to the left temporal lobe and to extra-striate and striate cortex (Gilbert and Li, 2013). In addition, one source of the neural generator of the N400, which reflects semantic access, has been shown to be in the temporal sulcus and the anterior temporal lobe (Halgren et al., 2002). These regions can easily feedback into nearby areas such as the parahippocampal gyrus and aid in visual processing. Thus, although the pathways for feedback from semantics into visual processing exist, we do not yet have all the details of how information flows between a subset of these regions, or the nature of any recurrent interactions within each region at these time scales.

A majority of the research in relation to semantics and visual processing has been done with isolated objects. In this work, I extend this knowledge into understanding the semantic

representations of natural scenes and their dynamic interaction with visual perception. Natural scenes are defined as photographs one would take of our surrounding environment without any alterations, for example a picture of a beach or a city street. They have high ecological validity as we live for the most part in these environments, yet they have not been as extensively studied as isolated objects. What we do know is that different natural scene categories co-activate visual regions, PPA and RSC (Epstein and Kanwisher, 1998), thus providing a more local specificity to the representations than studies using disparate objects (e.g. animals vs tools) that are known to activate disparate regions. In addition, we are well versed with what a typical natural scene for a category looks like (for example a beach). I will use both these aspects, the overlapping local co-activations of natural scenes in the brain and how well versed we are with natural scenes, to study the influence of semantics on the visual processing of natural scenes. In Chapter 2, I provide evidence that a common code exists for pictures and words describing natural scene images. This common representation is found not only in semantic processing regions but also in higher visual areas. Do these shared representations influence each other dynamically? In Chapter 3, to make contact with semantic representations, or prior knowledge, I use representativeness of natural scenes as an attribute to understand how and when does prior knowledge modulate visual processing. In Chapter 4, I use written words to precue categories and try and understand the dynamic interaction of semantics and the visual processing of natural scenes.

Chapter 2

Evidence For Similar Patterns of Neural Activity Elicited by Picture- and Word-based Representations of Natural Scenes[1]

2.1. Introduction

Seeing a furry, four legged animal with a wagging tail, hearing the sound of barking, and reading the word "dog" all evoke a (subjectively) common concept in our minds. What neural processes allow this common concept to emerge from processing that is initially modality and stimulus specific? A long-standing question is whether a common concept arises because these different stimuli all ultimately access the same representation – that is, elicit the same pattern of neural activity. In other words, is there a "common code" for semantic information in the brain that can be accessed from multiple modalities and stimulus types?

Before addressing the possibility of a common code, researchers needed to identify areas involved in representing conceptual information. Initially, univariate fMRI methods were used to find candidate brain regions important for conceptual/semantic processing. For example, researchers contrasted activity evoked by real words and pseudowords (which are perceptually like real words but lack learned semantics) or strings of consonants. This literature uncovered a distributed network of brain regions involved in semantic processing, including regions of lateral and ventral temporal cortex anterior to visual associative regions, the angular gyrus, the left inferior frontal gyrus, left dorso-medial prefrontal cortex, left ventro-medial prefrontal cortex,

[1] This chapter has been accepted for publication: Kumar et al., (in press), NeuroImage.

and the posterior cingulate gyrus (see review by Binder et al., 2009). Other studies have used pictorial stimuli (Chao et al., 1999; see review by Martin, 2007) and found a similarly distributed network of brain regions. A few studies have used paradigms with both types of stimuli to find brain areas that respond to both pictures and words when these are processed for semantics, getting slightly closer to the search for a common semantic code. Repetition suppression has been shown in the left fusiform region for both pictures and words (Kherif et al., 2010), and regions similar to the semantic network discussed above are activated by both pictures and words (Vandenberghe et al., 1996). Collectively, this work has revealed that there are a number of brain regions that are activated during the semantic analysis of words and pictures, and, based on these studies, some have proposed a model of semantic representation for concrete objects that is distributed across multiple regions, including sensory and motor systems (Martin, 2007; Pulvermuller and Fadiga, 2010).

Although these studies point to a distributed "common store" for semantic information, they are not sufficient to demonstrate the existence of a common semantic code. It is possible that the same brain areas become active when meaning is extracted from multiple input modalities and/or types, but that these brain regions nonetheless process each differently -- for example, using different subpopulations of neurons for each stimulus type. Thus, evidence for a common code in a particular region requires not only finding areas of common activation but also showing that different input modalities evoke similar representations, or shared patterns of activity, within those areas. To obtain this kind of evidence, the literature has turned to multi-voxel pattern analysis (MVPA; see review by Kaplan et al., 2015).

MVPA affords the ability to move beyond the extant univariate-based evidence supporting a common store by asking whether words and pictures evoke similar patterns of activation. For example, one can train a classifier on the pattern of activity from one type of stimulus (e.g., pictures) and attempt to then classify the pattern of activity elicited by a different stimulus type (e.g., words). We refer to this cross-modal training and testing of a classifier as "cross-decoding". Such studies have been performed using a variety of modalities (words and pictures: Fairhall & Caramazza, 2013; Shinkareva et al., 2011; pictures, written words, spoken words and natural sounds: Simanova et al., 2014). A few other studies have used Representational Similarity Analysis (RSA; Nili et al., 2014), a technique that uses distances between vectors (built from semantic feature lists or from the BOLD signal) to determine similarities between categories, in order to assess similarity in semantic representations across modalities (auditory words and pictures: Devereux et al., 2003; Liuzzi et al., 2015; written words and pictures: Bruffaerts et al., 2013). From these studies, cross-modal effects have been primarily detected in the left hemisphere: in the precuneus (lPrecu), posterior middle temporal gyrus (pMTG), inferior parietal sulcus (lIPS), precentral gyrus (lPCG), fusiform gyrus (lFG) and the inferior temporal gyrus (lITG).

The use of MVPA and RSA methods, then, have provided evidence that within the distributed semantic network, there are commonalities in the patterns of activation that are elicited by similar concepts across different forms of representation. There are two limitations of the extant literature on cross-modal representations utilizing MVPA methods. First, in comparison to MVPA studies in a single modality that explore fine-grained object categories (e.g. Kreigeskorte et al., 2007; Eger, Ashburner, Haynes, Dolan, & Rees, 2008; Borghesani et al., 2016), studies on

cross-modal representations have tended to use stimulus sets that varied across important semantic dimensions, such as animacy, size, and function (although see Bruffaerts et al., 2013 for a notable exception). Because of the substantive differences in their functional and motor affordances, some of these categories (e.g., tools and dwellings) activate clearly separate brain structures: e.g., dorsal motor regions in the case of tools versus ventral medial areas, such as the parahippocampal cortex, in the case of dwellings. In these cases, then, successful cross-decoding may reflect representational similarity at a fairly coarse level; that is, successful cross-decoding can reflect the fact that the objects activate very different regions of cortex. A more stringent measure of a common code in the brain would be to show representational similarity across categories that activate common brain. The few studies that have included fine-grained category distinctions (Fairhall and Caramzza, 2013; Bruffaerts et al., 2013) have been limited to individual objects as opposed to large-scale navigable natural scenes, which brings us to the second limitation of this literature. If we are interested in identifying cross-modal representations, such representations should extend beyond the object domain. Here, therefore, we sought to extend the cross-modal literature to natural scene categories, using four outdoor scene categories (beaches, cities, highways and mountains) known to activate very similar regions of cortex (Walther et al., 2009), making cross-decoding in those regions non trivial. Here, cross-decoding of category membership across representation type (pictures and words) must occur in a higher dimensional space (i.e. at a higher resolution) than for stimulus sets containing categories that show markedly different levels of activation across different brain regions. In other words, successful cross-decoding -- training classifiers on one modality and testing on another -- among these categories would necessarily imply locally similar neural patterns (i.e. within a restricted region of interest) between pictures and words.

Thus, in the present experiment, participants were scanned while they viewed full color photographs of real world scenes and, extending prior work that has mostly used single words (nouns), read two word phrases that described those categories of natural scenes (e.g. 'beautiful seashore'). By varying the specific noun that was used (e.g., seashore, beach, seaside) and pairing these with a range of adjectives (e.g., beautiful, humid, sandy), we provided a richer semantic stimulus while minimizing adaptation effects that might arise through simple repetition of just the category word (e.g. 'beach'). To test for evidence of a common semantic code (here, across pictures and words) and, more generally, to elucidate the semantic network involved in understanding natural scenes, we performed a cross-decoding analysis through the entire brain using a whole brain searchlight (Kriegeskorte et al., 2006). If we can successfully cross-decode from pictures to words and words to pictures, this would show that the category representations accessed from the two modalities is locally similar -- thus better supporting the existence of a common code.

2.2 Experimental Methods

2.2.1 Participants

Nine subjects (5 females and 4 males; two of the subjects were authors on the paper) participated in the study, which was approved by the Institutional Review Board of the University of Illinois. A tenth subject was dropped prior to analysis because his vision in the scanner had been uncorrected. All participants were in good health, with no past history of psychiatric or neurological diseases, and all gave their written, informed consent. The nine included subjects had normal or corrected-to-normal vision.

2.2.2 Visual stimuli and experimental design

Scene stimuli consisted of 64 distinct color images from each of four categories: beaches, cities, highways, and mountains, using images drawn from a similar set as Walther et al. (2009), which were downloaded from the Internet. Photographs were chosen to capture the high variability within each scene category.

Word stimuli consisted of 64 two-word phrases in each of the four categories. The first word in each phrase was an adjective and the second word was a noun (see Appendix A). Adjectives appropriate to each category were chosen. The adjectives were matched for word length and word log-frequency across all the categories using the celex database (Baayen et al., 1993). We chose three synonyms for the nouns in each category to make the phrases different and more engaging across the trials (beach category: beach, seaside and seashore; city category: city, town and downtown; highway category: highway, freeway and interstate; mountain category: mountain, peak and summit). The two-word phrases were unique within a category. Words were presented in a white font on a black background. Stimulus presentation and experimental design was controlled using the Psychophysical Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1998;) in MATLAB (Mathworks, Natick, Massachusetts).

Scene stimuli (800 x 600 pixels; subtending 14.3° x 10.4° of visual angle) or word stimuli (each letter was 1.04° in height and the word width ranged from 2.4° to 11.2°; presented using Arial font at 60 point size), were back projected on a screen viewed through a mirror from the bore of the scanner, using a projector operating at a refresh rate of 60 Hz. The experiment consisted of

16 runs alternating between the two conditions: the word condition consisting of the phrases and the picture condition consisting of images of scenes. Each run was comprised of eight blocks interleaved with 12 s fixation periods to allow the hemodynamic response to return to baseline levels. The first fixation period of each run was 20 seconds. Each category was presented twice in a run (4 categories x 2) and the order of categories was randomized across blocks, with the same order being preserved across pairs of word and picture runs. The starting run was counterbalanced between the two conditions, words and pictures, across subjects. Each block consisted of four presentations of pictures or word phrases from the same category.

In the picture condition, each scene image was displayed for 1 second with an inter-stimulus interval (ISI) of 1 second, for a total block length of 8 seconds (Figure 2.1A). In the word condition, each phrase was presented one word at a time (Note that, as this was a block design, subjects also knew the category of the noun when the adjectives were being presented). Each word was presented for a duration of 400 ms followed by a fixation period of 200 ms (Figure 2.1B). The total time for each phrase was 1000 ms, matching the time that the picture stimulus was displayed. An inter-stimulus interval of 1 second was maintained in the word runs, again matching that in the picture runs. The list of phrases for each category were grouped into sets of four (see Appendix A), and displayed for each block. Subjects always saw these sets of phrases as a block, with the sequence of phrases within a block being randomized across subjects. The presentation of these sets was randomized for each participant. A fixation cross was presented throughout each block, and subjects were instructed to maintain fixation. Subjects were instructed to view the pictures and read and think about the phrases silently. For one subject, the experiment was repeated at a later date due to a scanner error in the original session. Only data

from the second session was analyzed.

2.2.3 MRI acquisition and preprocessing

Imaging data were acquired with a 3 tesla Siemens Trio Scanner equipped for echo planar imaging. A gradient echo, echoplanar sequence was used to obtain functional images (volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 90°; matrix, 64 x 64 voxels; FOV, 190 mm; 29 axial 3 mm slices with a gap of 1 mm; inplane resolution, 2.97 x 2.97 mm). We collected a high resolution (1x1x1.2 mm voxels) structural scan (MPRAGE; TR, 1.9 s; TE, 2.25 ms, flip angle, 9°) in each scanning session to assist in registering our echo planar imaging images with localizer masks and a brain atlas.

2.2.4 Preprocessing of Imaging Data

We used the AFNI software suite (Cox, 1996) to pre-process our data. Motion correction was performed using the AFNI function 3dVolreg with options for zero padding (-zpad 4) and all volumes were registered to a volume in the 8th run. Our criteria were to reject any movement greater than 3mm and our subjects' head movements were under that threshold. The data were then normalized by subtracting the temporal mean of each run from the signal, dividing the result by the mean and multiplying the result by 100. This normalized signal was used for MVPA. No smoothing or any other pre-processing was done.

2.2.5 MVPA Searchlight Analysis

We performed a whole brain searchlight analysis (Kriegeskorte et al., 2006), using MVPA to build a classifier to discriminate stimuli category (beaches, cities, highways and mountains)

within a modality (words or pictures). In keeping with previous searchlight analyses of the scene network (Baldassano, Fei-Fei, & Beck, 2016), we defined a spherical template (radius of 7 mm), which contained 57 voxels (of size 2.97mm x 2.97mm x 3mm). These spherical voxel sets were built through the entire brain, with an 8mm center to center spacing. Voxels that fell outside the brain were omitted from the analysis. The fMRI data for these voxels, four data points per stimulus block, were extracted from the time series with a time lag of four seconds to approximate the lag in the hemodynamic response. In total we had acquired 256 brain volumes while viewing the pictures and word stimuli (1 session x 8 runs x 8 blocks x 4 images x 2 seconds presentation time/2 seconds TR). This extracted signal for each voxel set was fed as an input for pattern analysis.

A first step in the pattern analysis was to split the data into training and test sets. We followed leave-one-run-out cross validation (LORO) with 7 of the 8 runs being used as the training set and the left out run being used as the test set. We next scaled the data, which helps improve support vector machine (SVM) classifier performance (Chang and Lin, 2011), by normalizing the training data (7 runs) with respect to the mean and standard deviation of these seven runs. These same normalization parameters were used to normalize the data from the left out (8th) run. A SVM classifier was built (LIBSVM v3.11, linear kernel, C=0.01) using the training data set and trained to assign the correct scene category labels to the voxel activation patterns. The data from the test data set (8th run) was then presented to the trained classifier, which generated predictions of the class labels. In 8 repetitions of this procedure, for each condition (Words and Pictures), each of the 8 runs was left out once. For each repetition, the accuracy was calculated by counting the number of correctly classified stimuli for a category and dividing it by the total number of

actual stimuli for that category. Then, the mean of the accuracy across all the 8 repetitions was taken as the decoding accuracy for that voxel set for a condition (Word or Picture). The decoding accuracy for each template location was stored at the center voxel. By repeating this process for every voxel set, we obtained a brain mask of decoding accuracies for each subject.

To look for similarities across the group of subjects, we subjected the individual decoding maps to a group analysis. For the group analysis, we registered the decoding accuracy maps, converting them into 2mm x 2mm x 2mm maps for each subject into the Montreal Neurological Institute (MNI) space using the AFNI toolbox (Cox, 1996) @adwarp function. The @adwarp function transforms volumes from the subject space to the MNI space. We calculated the group mean using the AFNI 3dMean function across all subjects.

The pattern analysis procedure was slightly modified for cross-decoding. In this case we trained on one condition (e.g. Words) but tested on the other condition (e.g. Pictures). LORO cross-validation was again used, with training on 7 runs for one condition and testing on the corresponding 8th run for the other condition, and the decoding accuracy was determined as above.

2.2.6 Permutation Test

To determine an appropriate significance level for our classification results, we computed an expected distribution of classification errors by performing a permutation test (Mukherjee et al., 2003) with one thousand iterations, in which we randomized our category labels for the stimulus set and performed classification with these random labels. We executed this permutation test

separately for straight-decoding (pictures, words) and cross-decoding (words to pictures and pictures to words) for each subject. For each iteration of the permutation test, we calculated a group mean by transforming each subject's data into MNI space (see Group Analysis). Because the labels are randomly permuted, the 1000 resulting decoding scores per voxel provide a good estimate of the expected distribution of results under the null hypothesis for that voxel. To correct for multiple comparisons, from each of these 1000 group maps, we identified an accuracy threshold that resulted in clusters of voxels, for a given cluster size, in less than 5% of the maps. In this manner, we determined the minimum decoding accuracy associated with a $p<.05$ for each of the straight-decoding and cross-decoding analyses, for a total of four values. From these four values, to be conservative, we chose the highest cutoff value (i.e. 27.75% from the straight-decoding for words) that was considered significant at $p<0.05$, for a cluster size of 100 voxels. Our decoding and cross-decoding accuracy maps hence show decoding accuracies with a minimum threshold at 27.75%. We also performed a permutation test for a larger cluster size of 400 voxels: there, the cutoff accuracy for a $p<.05$ threshold was 27.0%. Similar results were obtained for a larger cluster size of 400 voxels, but we do see a few more areas in ventral temporal cortex at a threshold of 27% (see Appendix B). To adopt a more conservative approach, we report all results with the higher threshold -- a cluster size of 100 voxels and a minimum threshold of 27.75%, with the results for a cluster size of 400 voxels reported in Appendix A2.

2.2.7 Cross-decoding intersection maps

We created intersection maps of the two cross-decoding conditions (train on pictures, test on words; train on words, test on pictures) by intersecting the resulting maps from the group analysis. Our logic was that with respect to the idea that words and pictures produce a common

code, both of these directions of cross-decoding are equally valid. Regions that show above chance cross-decoding for both analyses thus represent stronger candidates for containing a common code.

2.2.8 Calculating Signal-to-Noise Ratio

As a preliminary analysis to assess our ability to detect an effect in the left Anterior Temporal Lobe (lATL), we computed the temporal signal to noise ratio (tSNR) in this region. In a region implicated in cross-modal representations (Binney et al., 2010) we created a 7mm sphere of voxels centered at coordinates (MNI: -39, -9, -36). This was transformed into each subject's space, and the tSNR (Friedman et al., 2006) was computed by calculating the temporal mean in each voxel and diving it by the standard deviation of the signal. The tSNR was also computed for other clusters that showed cross-decoding to serve as a comparison with the lATL, which is known to suffer from distortion and signal loss. We found, as expected, that the left ATL (MNI: -39, -9, -36) had the lowest tSNR (mean tSNR for picture runs = 42.92 and mean tSNR for phrase runs = 43.02) of all our cross-decoding regions (mean tSNR for picture runs range from 92.97 to 111.87; mean tSNR for phrase runs range from 96.93 to 114.24).

2.3. Results

Our goal was to assess whether there is representational similarity in the concepts evoked by pictures and words for natural scenes. To determine if pictures of natural scenes and phrases that evoke those scene categories share common representations, we used SVM to distinguish between four categories (beaches, cities, highways and mountains) under two conditions: straight-decoding or cross-decoding. Straight decoding, in which the classifier is trained and

tested on stimuli from one modality (e.g. pictures) will highlight regions that contain category information for a given modality. Cross-decoding, in which the classifier is trained on stimuli from one modality (e.g. pictures) but makes category predictions on a stimulus set from another modality (e.g. words), will then allow us to ask whether there are shared patterns of activation across the two modalities. If the different modalities share a common representation, cross-decoding should lead to above chance category prediction accuracies; that is, the pattern of activity evoked by pictures and words should be similar.

2.3.1 Representational Similarity: Cross-decoding

We used cross-decoding from pictures to words and words to pictures to test for representational similarity between pictures and words describing natural scenes across the entire brain. Because a priori we have no reason to value one direction of cross-decoding (e.g. train on pictures, test on words) over the other (e.g. train on words, test on pictures), we produced a combined map (Figure 2.2; Table 2.1) of both directions (an approach recommended by Kaplan et al., (2015)) by intersecting the individual maps (see Appendix D for cross-decoding maps in each direction). The resulting maps show regions with above chance cross-decoding both from pictures to words and words to pictures, ignoring regions that do not cross-decode in both directions (Figure 2.2). We see successful cross-decoding in a variety of regions, including putative visual areas -- the parahippocampal region, the retrosplenial complex (RSC) and the precuneus -- as well as regions attributed to semantic processing -- the angular gyrus, inferior frontal gyrus and the middle frontal gyrus. We note that successful decoding was not driven by any particular category; ROI analyses showed no significant differences across categories (see Appendix C for confusion matrices and BOLD signal data for regions with more than 100 voxels). A similar set of

distributed brain regions has also been seen in studies using pictures of objects and related nouns (Devereux et al., 2013; Simanova et al., 2014). Importantly, the successful cross-decoding in these regions implies that they contain a common code for semantic category across pictures and words.

Although we see above chance cross-decoding in large portions of the brain, we do not see significant cross-decoding in early visual cortex (Figure 2.2). Such a result, however, is to be expected given the visual features the stimuli (words vs pictures) share no commonalities. Interestingly, though, we do see cross decoding in later putative visual areas, the PPA and the RSC. Such data are consistent with recent findings indicating that semantic knowledge impacts visual detection (Caddigan et al., 2010; Greene et al., 2015). We also see cross-decoding in the caudal inferior parietal lobule (cIPL), a region we have recently argued might be part of a two-network model of scene perception (Baldassano, Esteva, Fei-Fei, & Beck, 2016). Not only is the cIPL functionally connected to scene processing regions (Baldassano, Beck, & Fei-Fei, 2013), but it, has been implicated in straight-decoding for natural scenes (Walther et al., 2009), processing familiar scenes (Montaldi, D., Spencer, T. J., Roberts, N., & Mayes, A. R., 2006) and in tasks involving spatial learning and landmark navigation (Bray, Arnold, Levy, & Iaria, 2015). It has also been implicated in semantic representations (Devereux et al., 2013; Binder et al., 2009) and has been argued to be a cross-modal hub (see review (Kravitz, Saleem, Baker, & Mishkin, 2011).

2.3.2 Decoding of Pictures Only

We also replicated previous results on decoding natural scene categories (Walther et al., 2009).

When training on picture stimuli and testing on untrained picture stimuli we saw above chance decoding of category (Figure 2.3A) bilaterally in: early visual cortex; higher visual regions in the medial temporal cortex including the parahippocampal place area (PPA) and the RSC; the lateral inferior temporal cortex; the inferior parietal lobule encompassing the precuneus, the posterior cingulate and the angular gyrus; the middle frontal gyrus; and the inferior frontal gyrus ($p < 0.05$, corrected). We also successfully decoded pictures in the perirhinal cortex, at the tip of the left anterior temporal lobe (ATL).

2.3.3 Decoding of Words Only

Two brain regions showed above chance decoding of category for words ($p < 0.05$, corrected): the left inferior frontal gyrus encompassing the pars Orbitalis, the pars Triangularis and pars Opercularis; and the left superior and inferior parietal lobule encompassing the precuneus (Figure 2.3B). These brain regions have also been reported as part of the semantic network from univariate studies using verbal stimuli (Binder et al., 2009). We successfully decoded in only a subset of the regions previously implicated in semantic processing, and the reduction in number of areas relative to pictures is notable. A similar reduction in decoding areas for words, as compared to pictures, has been observed in other MVPA studies for objects (Fairhall & Caramazza, 2013; Simanova et al., 2014). These studies show widespread straight-decoding for pictures in extrastriate visual areas, but word decoding only in the left pMTG/ITG. This is perhaps not surprising given that pictures are a much richer visual stimulus than words. One notable difference from univariate studies of words referring to objects (Kherif, Josse, & Price, 2010) is our failure to decode in the left lateral temporal cortex. We note, however, that natural scene categories are known to be represented in parahippocampal regions and not the lateral

temporal cortex (Epstein & Kanwisher, 1998), and thus we might not expect to see the same areas as those implicated in object semantics. But even if we instead consider only our scene-related areas, which do replicate prior work for natural scenes (Walther et al., 2009), we still do not see successful decoding of scene-related words in these areas. Of course, one cannot make too much of a failure to reject the null. Indeed, successful cross-decoding between words and pictures was observed in these regions, suggesting that the regions do differentiate concepts evoked by words.

Finally, it is interesting to note that the cross-decoding map is considerably more extensive than that produced by straight-decoding of words. This difference between cross-decoding and straight-decoding is also seen in other MVPA studies with words and pictures of objects (Fairhall & Caramazza, 2013; Simanova et al., 2014). At first glance this result may seem surprising; decoding is better between modalities than within a modality. However, such a result is possible in MVPA when the signal from one modality (i.e. words) is more vulnerable to noise, resulting in poorer decision boundaries during training and patterns that can on occasion erroneously cross decision boundaries during test. The addition of a stronger signal (i.e. pictures) not only allows the SVM to construct better (i.e. more generalizable) decision boundaries during training but it also provides a clearer signal with respect to those boundaries during test. Put succinctly, although the category signal is present for the words it does not always rise above the noise, but the presence of a stronger category signal from the pictures allows that relevant pattern to emerge more clearly.

2.4. Discussion

We successfully cross-decoded between pictures of natural scenes and phrases that describe them in a distributed set of temporal, parietal, and frontal brain regions. This cross-decoding implies a first-order similarity of neural representation across two different input types: The same patterns of activity that distinguish among visual scene categories distinguish among phrases describing the scene categories. Importantly, this cross-decoding was possible in small spatial windows, suggesting that the signals are locally similar across the pictures and words. Thus, our results go beyond showing simple overlap of processing in brain regions, and instead suggest the existence of a fine-grained common code in these regions. These multi-modal regions showing cross-decoding include higher visual areas such as parahippocampal gyrus and the reterosplenial cortex, regions in the parietal cortex, including the precuneus, the angular gyrus and the inferior parietal lobule, and the middle frontal and inferior frontal gyrus. Notably, all these regions have been implicated in semantic processing in other studies using a single input type -- for example words (see review Binder et al., 2009) or pictures (see review Martin, 2007). Our results suggest that words and pictures not only co-activate these regions, but actually activate them in a similar way.

We chose natural scene categories as our stimuli because they share many visual and semantic features and, not surprisingly, activate overlapping brain regions. Successful cross-decoding among these categories, then, must not only rely on more subtle differences among categories but it also must occur in higher dimensional space (i.e., at higher spatial resolution) than categorizations across stimuli that have less overlap in their activations. For example, dwellings and tools differentially activate large regions within the ventral visual cortex and premotor cortex (Chao et al., 1999; Martin & Chao, 2001). Such large differences in mean activation allow a

classifier to differentiate these categories in low dimensional space -- that is, on the basis of which area is more activated. The same is true for other objects classes that show other large-scale differences in activation, such as animate and non-animate object categories (Connolly et al., 2012). Thus, cross-decoding across words and pictures that include such object classes (Shinkareva et al., 2011; Simanova et al., 2014), does not imply a similarity of representation at as a high a spatial resolution as our natural scenes do. Successful cross-decoding of natural scenes in local regions thus brings us closer to a true measure of a common code.

2.4.1 Cross-decoding of natural scenes vs. objects

How do our cross-decoding results for scenes compare with those of prior studies using objects and nouns (Devereux et al., 2013; Fairhall & Caramazza, 2013; Simanova et al., 2014)? In the left hemisphere our cross-decoding regions were broadly similar to results for studies using objects (Devereux et al., 2013; Fairhall & Caramazza, 2013; Simanova et al., 2014). However, the cross-decoding regions in our study were more extensive in the right hemisphere than in previous work that used isolated objects (Devereux et al., 2013; Fairhall & Caramazza, 2013; Simanova et al., 2014). In particular, we found evidence for common semantic representations in the right angular gyrus and the right precuneus. Interestingly, these same regions are also seen in straight-decoding for pictures. Indeed, in our current study and in previous studies involving of natural scenes (Walther et al., 2009) we see more bilateral decoding of category for pictorial stimuli. Further work will be needed to examine the detailed properties that lead to differential engagement of the right hemisphere for isolated objects and natural scenes.

Although the anterior temporal lobes (ATL) have been shown with MVPA techniques (Clarke &

Tyler, 2014) to process semantic information, we failed to find cross-decoding there. This previous study used only picture stimuli, however, raising the possibility that this region has a unimodal visual representation. Some have argued that representations in the anterior temporal lobe are not domain-general, with the left temporal pole processing verbal semantics but the right temporal pole processing non-verbal semantics (Mesulam et al., 2013), and, indeed, other studies using words and pictures of objects also failed to find a common representation across modalities in the anterior temporal lobe (Bruffaerts et al., 2013; Fairhall & Caramazza, 2013; Liuzzi et al., 2015). However, on this account, we might still have expected to see bi-lateral straight-decoding in these areas, in the left for words and the right for pictures. Instead, we found significant straight-decoding of pictures in just a small region in the left ATL. One study, using only picture stimuli, has shown that this region, and more specifically the perirhinal cortex, is preferentially engaged when there is a need to process fine semantic distinctions (i.e. highly confusable objects: Clarke & Tyler, 2014). Although our scene categories overlap considerably in visual feature space, they are not highly confusable, and the perirhinal cortex therefore may have been less engaged in our experiment. A final consideration is that the poles of the ATL are particularly prone to fMRI susceptibility artifacts (Visser et al., 2010), and thus our failure to find significant decoding in those areas may simply be due to noise; we did not optimize our scanning protocol for the temporal poles. Indeed, tSNR in the ATL was the lowest of all the regions tested (see Experiment Methods, Section 2.8). In general, it is hard to draw conclusions about failures to find decoding. Thus, our main focus in this paper is to show that there are, indeed, some brain regions that do show cross-decoding at a high spatial resolution and therefore provide stronger evidence for a common semantic code.

2.4.2 Implications for models of semantic processing

A number of different accounts of semantic processing have been proffered in the literature (for review, see Glaser, 1992; Lambon Ralph & Patterson, 2008; Thompson-Schill, 2003). Multi-code models (reviewed in Barsalou et al., 2003; Thomspon-Schill, 2003) postulate that semantic information is represented as separate codes for each modality. For example, picture stimuli are represented by a completely different pattern of brain activity in a given brain region than are word stimuli that evoke the same concept. In conjunction with other recent work that has found evidence for shared representations of semantics from verbal and nonverbal material (Fairhall & Caramazza, 2013; Shinkareva et al., 2011; Simanova et al., 2014) -- i.e., evidence for a common code – our results argue against purely multi-code accounts, although they are not incompatible with the view that some aspects of semantic processing are modality-specific, as we find some brain regions that only exhibit straight decoding. Our results further provide evidence that there exists a common code at a higher spatial resolution than has typically been shown.

Common code models of semantic representation postulate that semantic information is represented in a form that is shared across modality and input type (Lambon Ralph & Patterson, 2008). Models of this type vary in whether they view the semantic system as consisting of a single amodal hub, for example the bilateral poles of the ATL, connected to a fronto-parietal semantic control system and also to modality specific sensory systems (Lambon Ralph & Patterson, 2008; Patterson et al., 2007), or distributed across brain regions (Martin, 2007) and in whether the shared code is taken to be fully abstract in nature (Lambon Ralph & Patterson, 2008; Patterson et al., 2007) or, at least in part, built from sensory/motor features that are nevertheless accessible from different types of inputs (Martin, 2007). Given that we find a distributed network

of brain regions in which cross-decoding is successful, our results are more consistent with models that posit that semantic representations arise from activity in a network of brain areas, as opposed to within a single (localized) hub. Moreover, given that some of the areas that show cross-decoding (such as the parahippocampal gyrus) are also generally taken to be part of sensory/motor processing networks, our results are consistent with accounts wherein perceptual representations form part of the semantic representation of concepts and are represented in the relevant perceptual system, using a common code in each.

Overall, then, the pattern of decoding results suggests that the processing of meaningful stimuli unfolds across a large-scale, distributed brain network, encompassing both sensory as well as cognitive processing regions (Figure 2.2). Parts of this network share a common code, while other regions perform analyses that provide high-level (i.e., categorical) information about specific input types. Moreover, it seems likely that the temporally-summed activity we measure here in fMRI reflects dynamic processing patterns that transition over time and brain area from modality- and input- specific processing to processing that is shared across input types (e.g., for a review see Federmeier et al., (2015)).

2.4.3 What is the nature of the representation in regions of cross-decoding?

Although our results show representational similarity across pictures and words for natural scene categories in multiple regions, we do not know the exact the nature of this representation. What is being decoded in these regions? Is it visual information, semantic information, or both? One possible view is that, as scenes and the words that describe them bear no visual resemblance to one another, any similarity of representation cannot be visual in nature and hence must be an

abstract, semantic representation evoked from a common concept (Fairhall & Caramazza, 2013). However, such a view overlooks the possibility that the semantics evoked by a word may integrally include visual perceptual representations that are stored in a visual feature space (Kan et al., 2003; Martin, 2007; Binder et al., 2005).

Considerations of visual imagery are particularly pertinent in visual processing regions. Indeed, visual imagery is known to evoke activity in the precuneus, the left angular gyrus, the supramarginal gyrus and the inferior parietal lobule (Ganis et al., 2004) – regions that show not only cross-decoding, but also straight-decoding for pictures and words in our study -- as well as posterior ventral visual cortex (e.g. O'Craven & Kanwisher, 2000), which also shows cross-decoding in our study. However, this imagery explanation does not need to be seen as an alternative to a semantic representation. An embodied view of cognition holds that the semantics of a concrete concept is distributed across both sensory and motor systems, which are then essential for the processing of semantic information (Martin, 2007; Pulvermüller & Fadiga, 2010). If every time someone reads the word "beach," for instance, his/her parahippocampal gyrus represents a beach or beach-like stimulus, then it is difficult to argue that this activity is not part of his/her concept of a beach. In other words, the activation of a visual representation, either through explicit imagery (D'Esposito et al., 1997) or while performing a semantic task (Kan et al., 2003), may be an essential part of semantics, at least for concrete (imageable) concepts. In the future, more work will be needed to understand how each of the regions in our cross-decoding network contributes to both our visual and semantic representations. It may be that the dichotomy that is sometimes assumed between semantic and visual information is at least partly artificial.

In summary, we have shown cross-decoding across picture and word input types for the natural scene categories of beaches, cities, highways and mountains. These stimuli, all constituting outdoor scene categories, evoke activity in similar brain regions. Thus, successful cross-decoding here implies a common representation at a high spatial resolution (i.e. the pattern is similar even within circumscribed brain regions). Although questions remain about the exact nature of the representation in each area, this commonality thus provides the strongest evidence to date of a common code.

2.5 Acknowledgements

The authors would like to thank the following people: Dr. Christopher Baldassano for his help with performing the permutation test; Mahmoud Elhrakhawy, Echo Ye, Keyur Kurani and Seema Dave for help with preparing the stimuli. Funding for this work was provided by the following: National Science Foundation IGERT Fellowship (MK), James S. McDonnell Foundation Scholar award (KDF), National Institutes of Health Grant 1 R01 EY019429 (LFF and DMB) and ONR MURI (LFF and DMB).

2.6 Table

Table 2.1. Location of various clusters from the cross-decoding map. The decoding accuracy at the peak location along with the co-ordinates are listed. We have added a suffix to distinguish two nearby clusters in the r-Middle Frontal Gyrus, the l-Middle Frontal Gyrus, and the l-Inferior Frontal Gyrus.

Clusters	Number of Voxels	Accuracy	Peak x	Peak y	Peak z
Precuneus, Angular Gyrus, RSC, IPL and SPL	11363	30.22	-28	-70	46
r-Middle Frontal Gyrus 1	458	29.97	26	6	64
r-Middle Frontal Gyrus 2	172	28.53	24	32	44
l-Middle Frontal Gyrus 1	124	28.53	-34	54	14
l-Inferior Frontal Gyrus 1	124	28.36	-36	34	12
l-Inferior Frontal Gyrus 2	73	28.32	-50	10	22
l-Middle Frontal Gyrus 2	56	28.62	-30	2	64
l-Paracentral Lobule	22	28.53	-8	-42	68

2.6 Figures

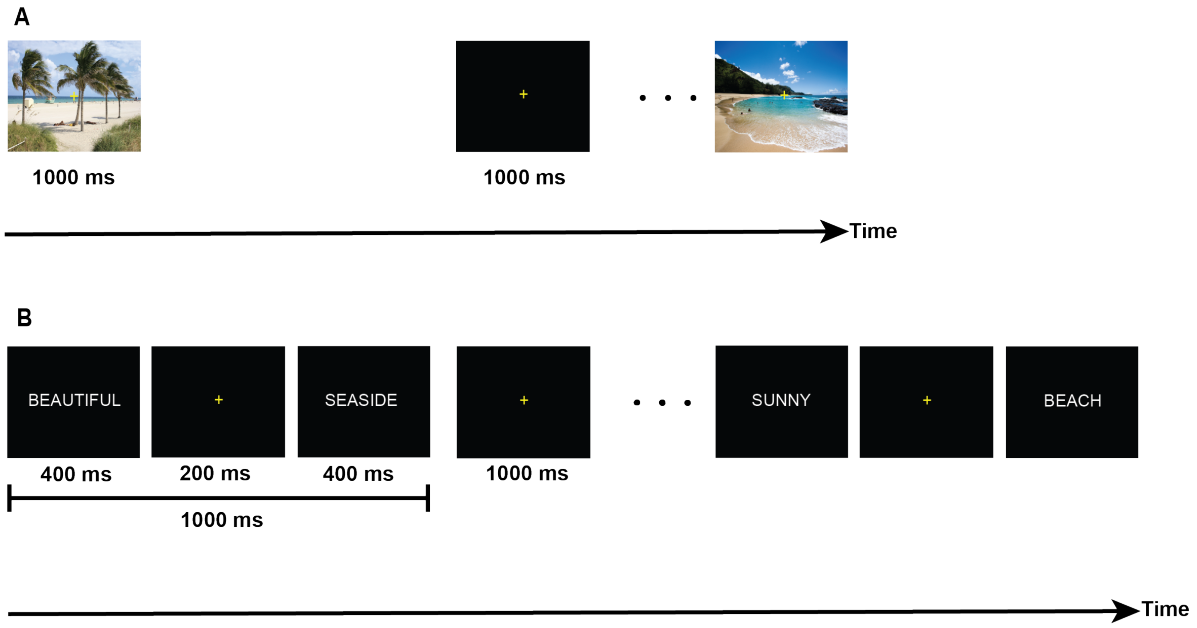


Figure 2.1. Stimuli were presented in separate picture (A) and word (B) runs. Presentations were blocked by category for both the picture and word runs. We used sixty-four unique stimuli in each of the four categories: beaches, mountains, cities and highways. Each image was presented for one second with an inter-stimulus interval (ISI) of one second. For the words, each stimulus consisted of two word phrases with each word being presented for 400 ms, with a fixation screen shown between words for 200 ms. Thus the total time for the word trial was also one second, equal to the time the pictures were shown. A one second fixation screen was shown at the end of the second word. Each block consisted of four trials for the picture and word runs.

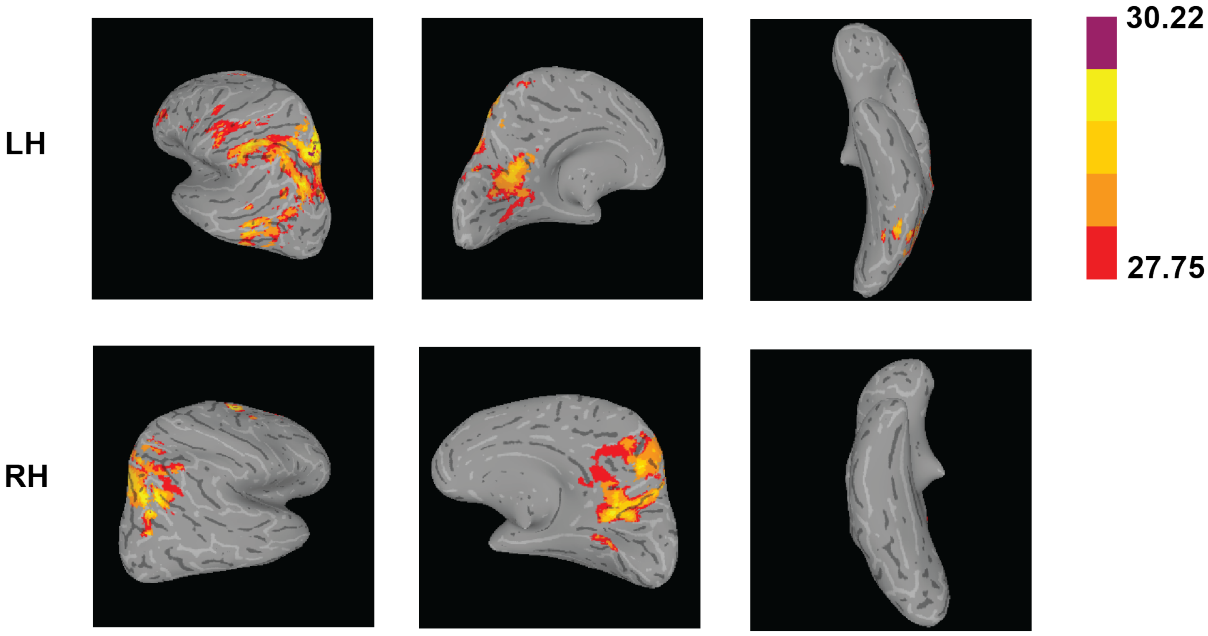


Figure 2.2 Maps showing the mean percentage of cross-decoding accuracy from pictures to words (train on picture runs and test on a word run) and from words to pictures (train on word runs and test on a picture run). The threshold for the maps have been set to an accuracy 27.75% ($p < 0.05$). The top row shows views from the left hemisphere and the bottom row shows views from the right hemisphere.

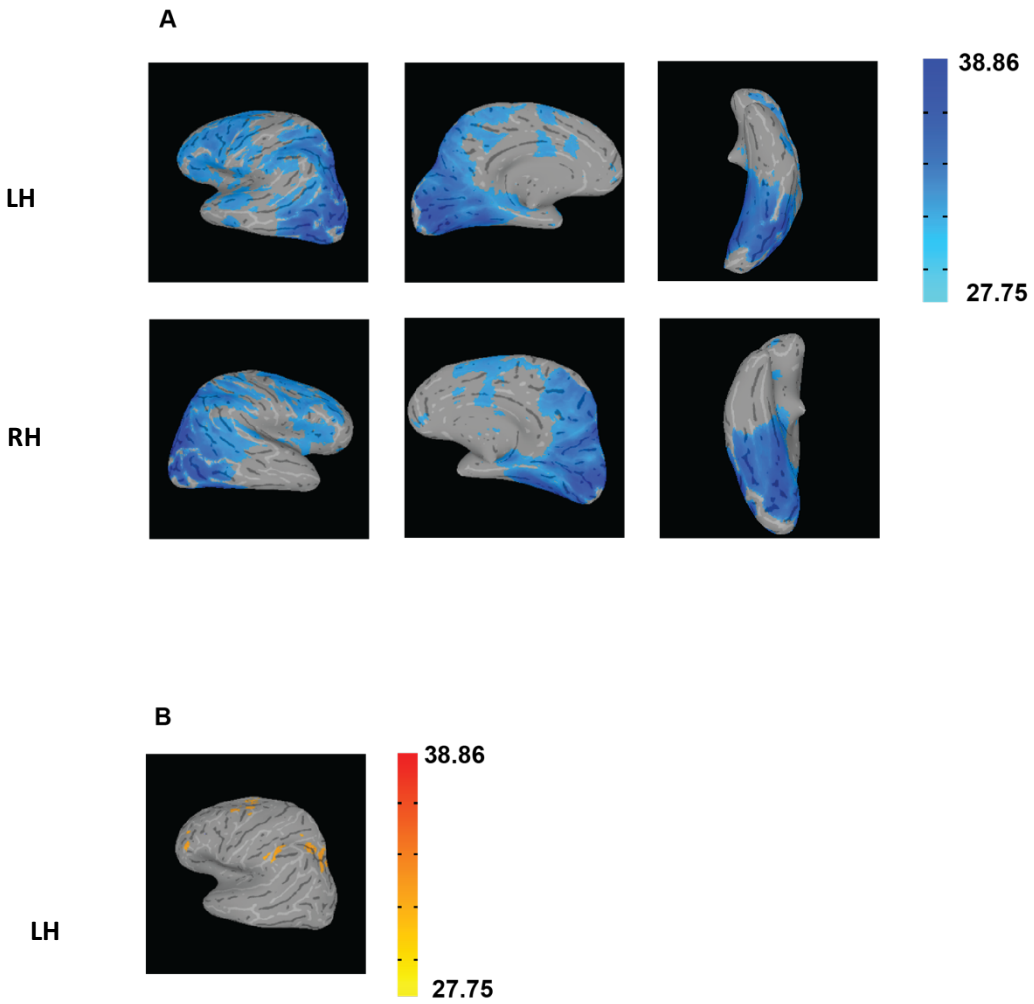


Figure 2.3. **A**: Straight decoding accuracy of pictures (train on picture runs and test on a left out picture run). The top row shows views from the left hemisphere and the bottom row shows views from the right hemisphere. **B:** Straight decoding accuracy of words (train on word runs and test on a left out word run). We see above chance straight-decoding for words only in the left hemisphere.

Chapter 3

The Good Bad Effect: How Does Representativeness Inform Vision?

3.1. Introduction

As a first step in understanding the nature of interactions between semantics and vision, we showed that there are similar representations across pictures and words pertaining to natural scenes, distributed across multiple brain regions, including higher visual areas and putative semantic brain regions (Chapter 2; Kumar et al., (in press)). Certainly, the early sensory areas, where we observed straight-decoding but not cross-decoding, seem to strongly represent modality specific information. Our cross-decoding results indicate that the higher associative areas, for example the PPA (an area known to process scene layout) and the RSC can activate representations across modalities. How do these modality specific features in early sensory areas and representations activated semantically in higher cortical regions interact when humans process natural scene images? When in the processing of natural scenes do we see an influence of prior knowledge or semantics?

One way to answer these questions is to compare scenes that make better contact with our prior knowledge, and hence have rich semantic associations, versus those that are less representative of their category. Representativeness of a scene image encompasses a wide array of visual properties that humans have learned about that category of scenes. By studying the processing of representative scenes (herein called *good*) as compared to non-representative scenes (herein

called *bad*), we can gain insights into when in the processing stream our prior knowledge about representativeness impacts visual processing

Good and bad exemplars of natural scenes have been studied using behavioral measures and fMRI (Caddigan et al., 2010; 2017; Torralbo et al., 2013). In a recent behavioral study (Caddigan et al., 2010; 2017), it was shown that subjects detect briefly presented good exemplars of natural scenes better than bad exemplars; that is, they can more quickly state that an image as opposed to noise was present when that image was a good exemplar of the category. Given that humans are good at extracting the gist of a scene, even at very fast presentation rates (Walther et al., 2009) and there is no gist in an image created with noise, we would not expect that detecting an intact scene from noise would depend on how representative the image is. Surprisingly, the representativeness of the scene matters. This result is a novel finding as prior work on categorization and detection does not inform us about the dependency of detection on categorization. We know that people are more accurate in categorizing good exemplars as compared to bad exemplars of objects (Rosch, Simpson, Miller, 1976). A study on detection and categorization of objects (Grill-Spector & Kanwisher, 2005) showed similar reaction times for categorization and detection, leading to the inference that categorization and detection occur in parallel. The better detection of good exemplars of natural scenes (Caddigan et al., 2010; 2017), thus extends the earlier work on categorization and detection, showing that even in a pure detection task, representativeness to a known category matters. We can think of good exemplars of a scene as being at the center of a multi-dimensional space that specifies the category, and hence evoke strong representations of the category. In the intact image versus noise detection

task (Caddigan et al., 2010; 2017), explicit categorization is not necessary and yet the degree to which the scene represents its category impacts perception.

Since our concept of a good beach, for example, must be learned, such an effect suggests that experience informs our vision. How is it that experience can have an effect on processing stimuli that are presented for just 10s of milliseconds and backward masked? Backward masking reduces the visibility of the target stimulus (Breitmeyer & Ogmen, 2000) by disrupting the ongoing processing of the stimulus with a mask that closely follows it in time. Nonetheless, subjects not only are able to perceive natural scenes even with very brief presentations (Walther et al., 2009; Li et al., 2002; Rousselet et al. 2002; Thorpe et al., 1996), they can perceive them with even less viewing time if they are good (as opposed to bad) exemplars of their category (Caddigan et al., 2010; 2017). There are both feedforward and feedback explanations of how representativeness impacts rapid perception. For example, exposure to natural scenes over one's lifetime may change visual cortex to enable quick feedforward processing of scenes and objects (Rousselet et al, 2005; Thorpe et al., 1996); that is, the superior performance on good exemplars does not initially rely on any semantic processing but arises through more efficient bottom-up processing of the visual features present. The other possibility is that even with incomplete perceptual processing, semantics/knowledge about scenes becomes active and feeds back into visual processing (Malcolm et al., 2014; Koivisto et al., 2011; Bar et al., 2007; Kveraga et al., 2007), making it easier/more robust. Our prior work (Kumar et al., in press; Chapter 2) is consistent with this possibility, given that we have shown that shared representations between verbal semantics and visual features exist in higher visual areas (albeit at much slower time-scales).

Prior work using fMRI (Torralbo et al., 2013) shows a range of possibilities at which good and bad exemplars can start to differ in their processing: they may differ in low level visual features, larger scale spatial layouts, or even at the semantic level. MVPA on data collected when participants viewed good and bad exemplars showed differences in decoding scene category between good and bad exemplars in brain regions spanning many different levels of analysis. For good exemplars, scene category could be decoded above chance in V1 as well as the PPA and RSC (Torralbo et al., 2013). Importantly, the decoding accuracy for good exemplars of the scene category was always higher than for bad exemplars of the scene category in all these regions. The decoding differences between good and bad exemplars in V1 imply that there could be differences at small spatial scales between good and bad exemplars. The decoding differences in the PPA and RSC imply that there could be differences between good and bad exemplars at larger spatials scales, as the PPA is known to process large spatial layouts (Epstein and Kanwisher, 1998). fMRI, lacking temporally specificity, also does not inform us about the time course of processing differences between good and bad exemplars. Thus, the fMRI results provide us a framework of possibilities that could contribute to the onset of differences between processing good versus bad exemplars, but the details pertaining to cognitive mechanisms and the time course of processing are lacking. Thus, we turn to ERPs to determine what cognitive mechanisms are responsible for differences in processing good and bad exemplars, and at what timescales do these processes differ.

Using ERPs we can distinguish between these possibilities based on the time-window in which differences in processing between good and bad exemplars are first observed. ERPs are a direct

and instantaneous measure of the continuous electrical activity in the brain (Münte, Urbach, Düzel, & Kutas, 2000; Luck and Kappenman, 2011) and can provide insights into the timeline and cognitive mechanisms underlying good and bad scene processing. Importantly, ERP signals are linked to cognitive processes spanning attention, perception, semantics, long-term memory and decision making (see overview Münte, Urbach, Düzel, & Kutas, 2000).

For example, if good and bad exemplar processing differed due to low level visual features, we would predict effects on sensory components such as the P1 and N1 in the time-window 100-150 ms (Schechter et al., 2005). If the differences between processing good and bad exemplars were due to differences in the semantics evoked, we would see differences on the N400 component in the timeframe of 350-600 ms (Federmeier and Kutas, 2001). If differences between good and bad exemplars are due to a judgment and decision making stage we would see differences in the late positive complex (LPC) component after about 500 ms (Finnigan, Humphreys, Dennis, & Geffen, 2002). If differences between good and bad exemplars arise due to the interaction of prior knowledge and visual processing (Caddigan et al., 2010; 2017), it is plausible that we will see those differences in the N300 component (Schendan and Kutas, 2002) which indexes higher level perceptual processing in time-window 200-350 ms, since this is a time-frame in which both semantics (350-600ms) and high level perceptual processing could be active (200-350).

We used good and bad exemplars from six natural scene categories: beaches, forests, mountains, city streets, highways, and offices to examine the cognitive and perceptual processes underlying differences between good and bad exemplars. We measured ERPs to understand the mechanisms

that lead to differences between processing good versus bad exemplars and the time when these differences occur.

3.2 Methods

3.2.1 Participants

Twenty right-handed neuro-typical college-age subjects (mean age = 23.9 years, range = 18 to 33 years) participated in the study. Participants signed an informed consent and were compensated for their participation in the study, through course credit or monetary compensation. The study was approved by the Institutional Review Board of the University of Illinois at Urbana-Champaign. All participants were right-handed, as assessed by the Edinburgh Inventory (Oldfield, 1971). Participants also had no history of neurological disease, psychiatric disorders, or brain damage.

3.2.2 Visual Stimuli

We choose pictures of natural scenes in six categories: beaches, forests, mountains, city streets, highways and offices. These images were collected from the internet and rated for representativeness to category on Amazon Mechanical Turk, and the top rated 60 images were marked as good exemplars for each category, and the lowest rated 60 images were marked as bad exemplars for each category (for details on the choice of good and bad exemplars see Methods in (Torralbo et al., 2013). Using an image processing software (Imagemagic, http://imagemagick.org/script/index.php), these images were resized to 340 x 255 pixels and placed on a black background with a fixation cross placed at the center. The images were randomly presented, for each trial, at one of three locations: the center, placed 2 degrees to the

left of fixation, or 2 degrees to right of fixation, with a total of 120 good images and 120 bad images presented at each location. In total 360 good images and 360 bad images were presented across the three locations. In this work, we only report results from the images that were centrally placed. The results for the lateralized images will be reported as a part of a study on hemispheric differences in the future.

3.2.3 Presentation

Subjects were instructed at the beginning of the study that they would be seeing good and bad exemplars of six scene categories and that their task at the end of each trial was to indicate via button press whether the image was a good or a bad exemplar its category. They were given a few practice trials to get them accustomed to the task. They were seated at a distance of 100 cm from the screen, and the images subtended a visual angle of 7.65° x 5.73° (width x height). Subjects were instructed to fixate on a central cross. They were told to remain relaxed throughout the experiment and that they could blink their eyes once the trial was complete, before they hit the response button. Responding would start the next trial. Each trial began with a fixation cross presented on a blank screen for a duration jittered (to prevent any expectancy effects) between 1000-2000 seconds (Figure 3.1). This was followed by the presentation of an image, either a good exemplar or a bad exemplar from one of the six categories, for a duration of 200 ms. This was followed by a fixation cross on a blank screen for 500 ms. At the end of the trial a prompt with "Good or Bad?" was displayed on the screen and subjects used a button press to indicate their judgment. The experiment lasted for approximately one hour and fifteen minutes. Subjects were given two five minute breaks at roughly 25 minutes and 60 minutes from the start of the experiment.

3.2.4 ERP Setup and Analysis

We used 26 channels of passive electrodes that were equidistantly arranged on the scalp using an electrode cap. In addition, we used 3 electrodes on the face to measure eye movements and blinks. Impedances were kept below 5 KΩ for scalp channels and 10 KΩ for eye channels. Horizontal eye movements were tracked by computing the difference between signals extracted from electrodes placed on the outer canthus of the left and right eye. Eye blinks were tracked by placing one electrode below the left eye. The signal was bandpass filtered online (0.02 Hz - 100 Hz) and sampled at the rate of 250 Hz. The EEG signals were converted into voltage from their analog to digital (A/D) values by independently calibrating the A/D units of the amplifier with preset voltages. Artifact rejection was performed before averaging the signal to remove excessive eye movements, drift, and blinks, using thresholds calibrated for individual subjects A blink correction algorithm (Dale, 1994) was used on all subjects to recover signals due to blink correction. The EEG signals were re-referenced offline to the mean of the left and right mastoids. ERPs were calculated for a time period spanning 100 ms before stimulus onset to 920 ms after stimulus onset, with the 100 ms prestimulus interval used as the baseline. This processed signal was then averaged for each condition across all subjects. A digital bandpass filter (0.2 Hz - 30 Hz) was applied before measurements were taken from the ERPs.

We computed the grand average ERP waveforms for good and bad conditions across all subjects. For the two conditions, the mean amplitude was computed at frontal electrode sites and t-tests and bayes factors were computed. We also computed the grand average ERP waveforms for natural (beaches, forests, and mountains) and man-made (city streets, highways, and offices)

categories. We computed statistics using R (R Core Team, 2014). Specifically, we used the functions t.test, to compute t-tests, and ttestbf (from the package: BayesFactor) to compute Bayes Factors. For within-subject calculations of confidence intervals, we used the function summarySEwithin() that is based on Morey (2008).

3.3 Results

The grand-averaged mean ERP signal for each condition is shown at 8 electrode sites in Figure 3.2A. This figure shows that the good and bad exemplars appear to be differentiated beginning around 250 ms from stimulus onset, with greater negativity for bad exemplars than for good exemplars. The timing, polarity, and scalp distribution is consistent with the N300 (McPherson & Holcomb, 1999; Schendan and Kutas, 2002; 2003; 2007). To measure this N300 response, which is known from prior literature to be frontally distributed (REFs), we computed mean amplitudes for good and bad exemplars from the 11 frontal electrode sites. As shown in Table 3.1 and Figure 3.2B, this analysis revealed that bad exemplars elicit significantly larger (more negative) N300 responses than do good exemplars.

We also see differences between good and bad exemplars in a 500-800ms time window at parietal sites, with greater positivity for good exemplars than bad exemplars $F(1,19) = 12.61$, $p=0.0021$, epsilon = 0.0845 (Greenhouse-Geiser). The ERP waveforms in this window show the form typically described for the Late Positive Component (LPC), which has been associated with confidence in decision making (Finnigan, Humphreys, Dennis, & Geffen, 2002). This effect suggests that subjects had greater confidence in judging the good exemplars as good, as opposed to the bad exemplars as bad. Indeed, subjects were better at identifying good exemplars

(accuracy: Mean =85.83%, std. dev = 14.17%) as compared to bad exemplars (accuracy: Mean =55.83%, std. dev =15.83%).

As a point of reference to prior work, we also examined the differences in ERP waveforms between natural and man-made scene categories. In categorization studies, the natural vs man-made is an easy distinction to be made on the basis of features (Oliva & Torralba, 2001), and is made before basic-level categorization (Loschky & Larso, 2010) and thus is a good candidate for differing in low-level processing. Indeed, we do find that the ERP waveforms begin to differ for natural scene categories as compared to man-made scene categories approximately 150 ms from stimulus onset (Figure 3.3). This is interesting because it suggests the good/bad distinction is more complex than the natural/man-made distinction since good/bad distinction comes considerably later.

Table 3.1. The grand average mean values, in the N300 time-window (250-350 ms), shown for 11 frontal electrode sites (see Figure 3.2B) along with t-test and Bayes factor values. The N300 for bad exemplars have a greater mean negative amplitude than for the good exemplars. The t-test and Bayes factor calculations compared the Good/Bad difference to 0.

Condition	N	Mean (micro Volts)	Standard Deviation (micro Volts)	t(19)	p	Bayes Factor
Bad	20	-6.4	0.7	-5.35	3.64E-05	686.7
Good	20	-5.3	0.7			

3.4 Discussion

We set out to answer one question: What are the mechanisms that the human brain uses to differentiate between good and bad exemplars of natural scene images. Thus, we used ERPs, from which we can not only get timing information but also gain insights into the perceptual and cognitive processes that help distinguish good and bad exemplars. We found that the ERP waveforms for good and bad exemplars differ in processing in the time-window of 250-350 ms. These waveforms show a greater negativity for bad exemplars than for good exemplars and this effect is distributed across frontal electrode sites. Given the time window of the difference between the ERPs for good and bad exemplars in our experiment, and their frontal distribution, we characterize these waveforms as reflecting the N300 component, which has been characterized in other work using images of objects (Schendan and Kutas, 2002; 2003; 2007) and scenes (Pietrowsky et al., 1996; Vo and Wolfe, 2013).

The N300 has been shown to index the degree of structure and regularity in pictorial stimuli. In a study with pictures of objects at different levels of fragmentation (ranging from a fuzzy collection of line segments to an arrangement of line segments where the object outline is very clear), the ERPs in the time window of 200-350 ms showed greater negativity, over frontal sites, when subjects were unable to identify the object segments as compared to when they did recognize the object (Schendan and Kutas, 2002). The N300 is not sensitive to local differences in contours: when low level properties (e.g. color of the line segments) are changed, there is no difference in the N300 time window for the original and changed versions of the stimuli (Schendan and Kutas, 2007). The N300 is also sensitive to well learned viewpoints of objects. Non-canonical views (e.g. An umbrella held horizontally) elicit a greater frontal negativity as compared to a canonical view (e.g. An umbrella held vertically) (Schendan and Kutas, 2003).

Similarly, in a study using natural scenes, the ERPs were more negative in the N300 time window when the scene was scrambled (created by recombining parts of the scene image into a random jigsaw) as compared to when the landscape scene was intact and identified (Pietrowsky et al., 1996). This negativity in the N300 time-window was also seen in a study with natural scene images that contained elements not in their probable locations (e.g. A cup placed on top of a microwave door) as compared to a picture of scene where all elements were in predictable places (e.g. A cup placed inside a microwave) (Vo and Wolfe, 2013). Thus, the N300 component helps distinguish across a variety of perceptual contexts that do not just differ in low level attributes but encompass global structure, canonical viewpoints, probable views of scenes, and in our own experiment, the representativeness of the stimuli.

We would like to collectively refer to these properties -- canonical viewpoints, probable views and representativeness – as learned statistical regularities of the stimuli. It is this collection of statistical regularities, which we can also call a *template*, that can help us in rapid categorization and identification of stimuli. Thus, we can think of the differences in N300 component as an indicator of the degree to which the exemplars match a template, with greater negativity for a stimulus when it doesn't match a template as compared to when it does match a template.

These templates play an important role in object and scene recognition. An object recognition model has been proposed, based on a series of experiments characterizing the N300, to occur in multiple stages for familiar and unfamiliar items (Schendan and Kutas, 2002; 2003; 2007). In the first stage, indexed by the P150-N170 components, perceptual grouping processes are at work. This stage constructs the whole from parts and helps identify easily separable objects (faces vs.

non-faces) or even indicates an early match to a canonical viewpont. In the next stage, called the object selection stage, indexed by the N300, objects are matched to items in memory with similar perceptual structures. This stage is viewpoint centric, as indexed by the N300, and reflects the processing of the global structure of the object. Thus, at this stage, canonical views and previously seen objects are identified (lower N300 amplitude). If sufficient information is available to identify non-canonical items, they can be identified at this stage too, indexed by a larger N300 amplitude, as compared to the canonical views. This stage involves recurrent and feedback processing (David, Harrison, & Friston, 2005; Schendan and Kutas, 2007) and results in object selection for canonical and identifiable objects. The N300 stage of processing has also been proposed as a pruning stage, where representations matched to perceptual structure in memory are kept and all other representatons pruned out from selection (Pietrowsky et al., 1996). For stimuli that have no identifiable information (non-recoverable), object selection is not made at this stage, and no modulations are seen in the N300 amplitude. For these non-recoverable stimuli, a later stage identification process manifests after 500 ms, indexed by the LPC. At this stage, identified objects, and hence matched to long-term memory, show a greater positivity as compared to unidentifiable objects. This LPC stage is also where fine distinctions, at local scales (e.g. positions of small edge segments), are made between objects (Schendan and Kutas, 2007).

Our work extends this understanding of the N300 and the cognitive and physiological processes that it indexes, as there are some notable differences between our bad exemplars and the stimuli used in all the previous work characterizing the N300 component. Our bad exemplars are not impoverished images of isolated objects, unlike the fragmented stimuli used in prior studies, nor do they have any artificially displaced elements in them. Rather, they are full color photographs

of natural scenes that don't match our expectations about representative views of the category. In the object model (Schendan and Kutas, 2002; 2003; 2007), objects were matched to known or canonical viewpoints and showed a lower N300 as compared to objects that had non-canonical viewpoints. The mismatch to regularities, or templates, is the key insight that we can transfer to the interpretation of our study from prior studies on N300 differences. Good exemplars match a learned template better, and hence show a lower N300 amplitude, as compared to bad exemplars.

We can also draw parallels between the object identification models (Schendan and Kutas, 2002; 2003; 2007). and scene categorization models (Oliva and Torralba, 2001). Natural scene categorization can be achieved at low spatial resolution, that is without any identifying information about objects in the scene (Oliva & Schyns, 1995). One model, the Spatial Envelope model, posits the use of these spectral signatures in extracting the gist of a scene (Oliva and Torralba, 2001). Gist extraction is from the global image (Greene & Oliva, 2009) and hence fits the time-window of the N300, based on the object studies (Schendan and Kutas, 2002; 2003; 2007). The templates we propose, constituting learned statistical regularities, can serve to select and match gist in a perceptual structure system, with images that find a template match are facilitated and identified in the N300 time-window. In our case, all our images are complete, as in not fragmented or non-recoverable. Hence, most of the processing and identification can happen at the N300 stage, similar to the object model (Schendan and Kutas, 2002; 2003; 2007). Scene processing is known to occur in extrastriate visual areas in the PPA and RSC (Epstein and Kanwisher, 1998) corresponding well with the N300 localization in higher visual areas (Schendan and Kutas, 2002). Similarly, recurrent and feedback processing, as in the object

model (David, Harrison, & Friston, 2005; Schendan and Kutas, 2007), do play a role in scene processing (Malcolm et al., 2014; Serre et al., 2007).

We have thus gained new insights into how our prior knowledge, stored as templates, of natural scenes, is used in scene processing. In the next chapter, we focus on examining the dynamics of processing good and bad exemplars of natural scenes under conditions in which participants are precued, and hence can pre-activate features relevant to the incoming stimuli category template. This will help address the question: Is a good exemplar always facilitated in processing or does what we expect modulate the processing of good exemplars.

3.5 Figures

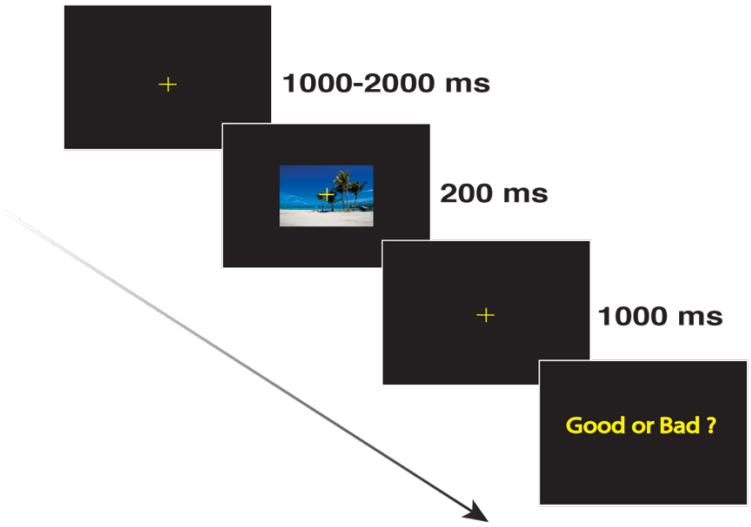
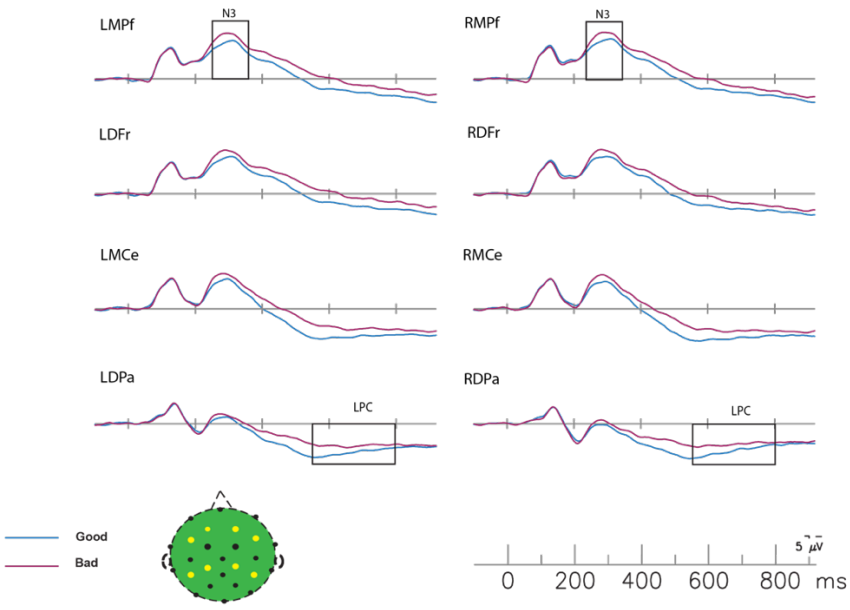


Figure 3.1. Schematic of one trial in the experiment. A screen with a fixation cross is shown for time-period randomized between 1000-2000 ms. A good or bad exemplar image from one of the six categories is presented at one of the following locations: the center, the right visual field or the left visual field, followed by a fixation cross. The subjects then make a delayed response to the question "Good or Bad?" and the next trial begins.

Figure 3.2A


LMPf
N3
RMPf
N3
LDFr
RDFr
LMCe
RMCe
LDPa
LPC
RDPa
LPC
Good
Bad
5 μV
0
200
400
600
800
ms

B.

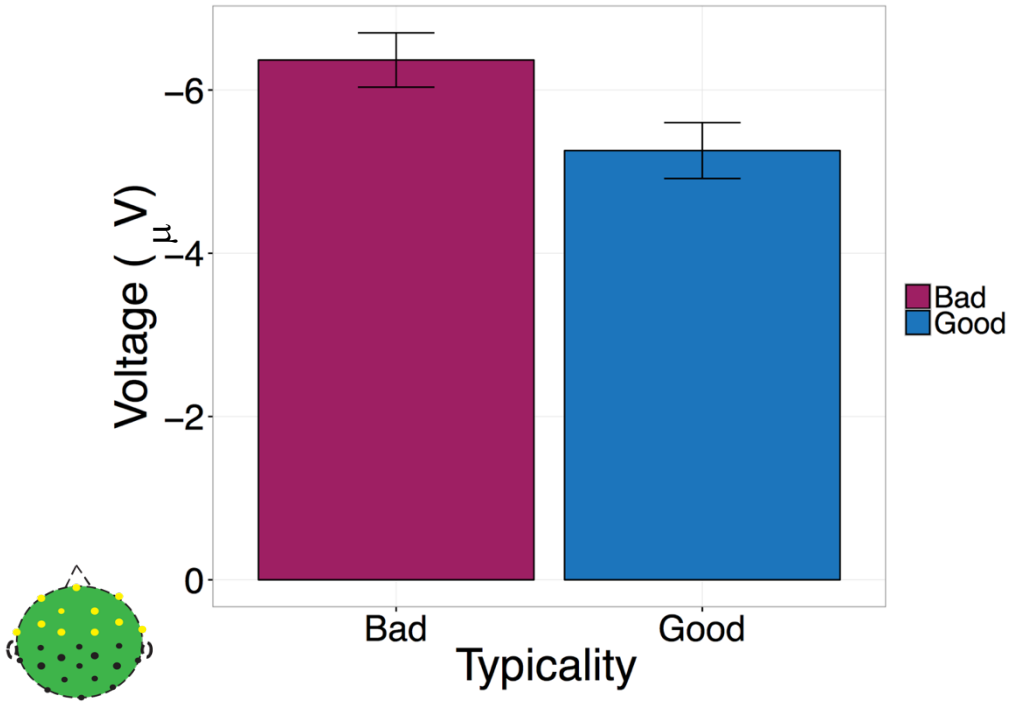

Voltage (μV)
−6
−4
−2
0
Bad
Good
Typicality
Bad
Good

Figure 3.2 A (contd.): The grand average ERP waveforms for good (blue) and bad (maroon) exemplars shown at 8 electrode sites. Negative voltage is plotted upwards. The channel locations are marked in yellow on the schematic of the scalp. The waveforms differ in the N300 time-window (250-350 ms), with greater negativity for bad exemplars as compared to good exemplars, over frontal sites. **B**: The mean of the ERP amplitude over 11 frontal electrode sites (N = 20). The error bars plotted are within-subject confidence intervals.

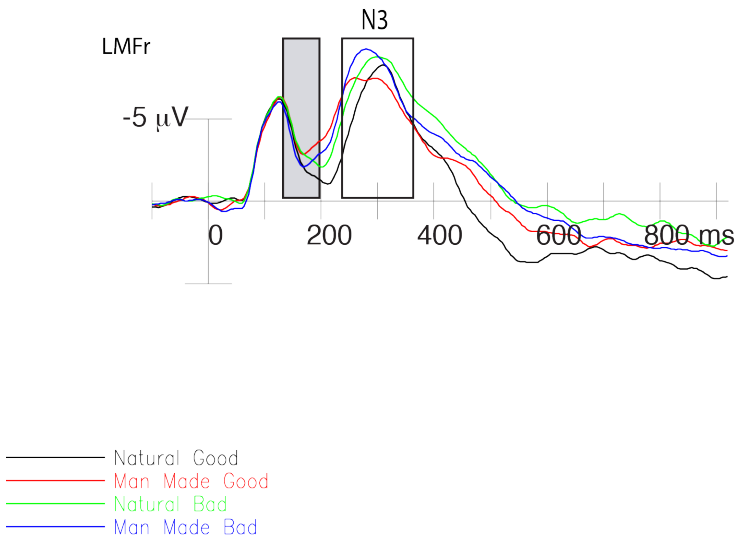


Figure 3.3. The grand average ERP waveforms plotted for natural and man-made scenes for good and bad exemplars. The natural vs. man-made distinction is seen separating in an earlier time-window, approximately 150-200 ms after stimulus onset in all fronto-central channels, as compared to the separation for good versus bad exemplars. The grey box indicates the time-window of the separation between the natural and man-made categories, for both good and bad exemplars, showing that is earlier than the time-point for the good versus bad distinction.

Chapter 4

Expectancy Modulates the Good-Bad Effect

4.1. Introduction

I have showed, so far (Chapter 3), that the difference in processing between good and bad exemplars occurs in a time-window of 250-350 ms, with the bad exemplars showing a greater negativity in the ERP waveform over frontal electrode sites. I linked this processing to the ERP N300 component that indexes global structure, canonical viewpoints, and – now – representativeness to category. I collectively refer to these properties as learned statistical regularities, or "templates" for short. I hypothesized that good exemplars are facilitated in their processing because they better match a template for the category as compared to bad exemplars. In this chapter, then, I focus on the dynamics with which those templates are called to mind and used. Specifically, I ask what happens to the visual processing of incoming good and bad exemplars of a category if people dynamically modulate their expectations about stimulus category through the use of a prior, verbal cue. Does this dynamic activation of semantics, via reading a word cue, influence picture processing in the N300 time-window (or earlier) and thus modulate how statistical regularities are accessed and used during visual processing. The answer to this question will give us insights into the dynamics of the interaction between semantic and visual processing.

Building on the design from Chapter 3, in this experiment we set up an expectation for a particular scene category and then present either a good or bad exemplar of the cued scene

category or a good or bad exemplar of a mismatching scene category. A number of prior studies have examined basic semantic match/mismatch effects ; (see review Kutas and Federmeier, 2011), including from verbal cues to images, allowing clear predictions about the type of effects that should be seen in later processing windows. In particular, semantic mismatch effects (e.g. reading the word "Forest" followed by a picture of a beach) would be expected to modulate the amplitude of the N400 component. The timing of the N400 for a stimulus (words, sentences, pictures, sounds, gestures) is stable, seen between about 300 and 500 ms, peaking at around 400 ms (Kutas and Hillyard, 1980; see review Kutas and Federmeier, 2011). The N400 component is distributed over central-parietal sites for words and fronto-central sites for colored pictures. This is true not only for written sentences but also for prime-target word pairs (e.g. congruent*: farm-ranch*; incongruent: *hook-table*), with the largest N400 amplitude seen for target words that are outside of the category of the first word (Kutas, 1993). This prime-target paradigm is what we are using in our current study, except the target is a picture of a natural scene. Given prior work, we expect the N400 pattern to be similar for pictures. A similar pattern of N400 results was obtained for written sentences with the last words sometimes replaced with a picture (Federmeier and Kutas, 2001; Nigam et al., 1992); for spoken sentences with pictures (congruent or incongruent) simultaneously displayed when the critical noun (congruent or incongruent) (Willems, Özyürek, & Hagoort, 2008); for line drawings (Holcomb & Mcpherson, 1994); for colored pictures (Mcpherson & Holcomb, 1999); and for video clips (Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008). Thus, in our experiment, if the scene stimulus (e.g. picture of a beach) matches the verbal cue category (e.g. Beach) we expect to see reductions in the N400 amplitude to the match as opposed to when the stimulus (e.g. picture of a forest) does not match

the verbal cue (e.g. Mountain) and hence are not in the semantic neighborhood of the cue, leading to large N400 amplitudes.

What is less clear from the extant literature is what the impact of good versus bad might be on the N400 pattern. If what makes a scene good (e.g. a good beach) is that it is better at evoking stored semantic associations for the category (e.g. beach), we expect to see the strongest N400 effects for good exemplars, with robust reductions in N400 amplitude to good exemplars that match the cue (e.g. Beach) versus good exemplars that do not match the cue (e.g. Mountain) and hence are not in the semantic neighborhood of the cue. On this view, we might expect an intermediate response to matching bad exemplars, which share some of the category's features but not as many as the matching good exemplars. Mismatching bad exemplars are an interesting case. If they share no features with the cue, then N400 amplitudes to mismatching items should be insensitive to the good/bad dimension – that is, both conditions should be at the baseline (large N400), unprimed level. However, if part of what make bad exemplars bad is that they share some semantic features with other scene categories, then N400 responses to this condition might be facilitated relative to the good mismatches.

Of greatest interest for the present study, however, is the effect on the N300. If the N300 reflects a point in processing in which knowledge-based information interacts with vision, it should be possible to dynamically influence the process by setting up knowledge-based expectations? By setting up an expectation, we are plausibly activating a template of statistical regularities of the category that is expected. That, in turn, may modulate the good/bad effect since, e.g., an incoming stimulus that is a good exemplar of beaches would still be a poor match for an

activated template for forests. Prior work and our own results indicated that the N300 time-window is when object selection and matching (Schendan and Kutas, 2002; 2003; 2007), pruning (Pietrowsky et al., 1996), or matching to templates occurs (Chapter 3). Thus, we are asking whether the matching process is sensitive to an active template, initiated from a different input type (written word). If the N300 component is sensitive to the active template, set up before the stimulus, then the arriving beach stimulus is not a match and we should see a larger N300 for the good exemplar of a beach as compared to when the cue (e.g. beach) and the incoming stimulus match (e.g. picture of a good beach) -- that is an unexpected good exemplar is processed similarly to a bad exemplar (i.e. a less good match to the active template) in the N300 time window. If the N300 component is indexing templates activated only by the incoming stimulus and cannot be modulated by any preexisting activations, we should see similar facilitation for good exemplars regardless of whether it is preceded by a matched or mismatched cue.

In this work I use a precue to assess the dynamics of top-down semantic influences on the visual processing of natural scenes. Similar to the previous experiment (Chapter 3), I presented images of good and bad exemplars from the same six categories, but with one change: I cued participants with a noun describing the category before the start of the trial, thus setting up their expectation of what they would see. For seventy five percent of the trials, the cue matched the stimulus, constituting either a good or a bad exemplar from the cued category, but for twenty five percent of the trials, there was a mismatch between the cue and the image (Figure 4.1) (See Methods 4.2 for more details).

4.2 Methods

The methods were almost identical to those in the previous chapter (see Methods Section 3.2), with the following differences:

4.2.1 Participants

A different set of twenty-four subjects participated in this experiment. We had to drop four subjects in total due to excessive eye-movements and noise (3 subjects), and problems with data recording for one subject. (mean age = 23.9 years, range = 18 to 33 years). Participants signed an informed consent and were compensated for their participation in the study, through monetary compensation. The study was approved by the Institutional Review Board of the University of Illinois at Urbana-Champaign. All participants were right-handed, as assessed by the Edinburgh Inventory (Oldfield, 1971). Participants also had no history of neurological disease, psychiatric disorders, or brain damage.

4.2.2 Presentation

The trials began with a word cue, presented for 500 ms (Figure 4.1). The cue consisted of a category cue from one of the following words: Beach, City Street, Forest, Highway, Mountain, and Office. For each category, we ensured that five trials were mismatched, for good and bad exemplars. We thus had 75% matched trials and 25% mismatched trials. At the end of each trial, participants were shown a question,“Yes or No?” Participants had to make a delayed button response and indicate if the picture matched the cue or not. The responses for “Yes” and “No” were counterbalanced to the left and right hand across subjects.

4.2.3 ERP Setup and Analysis

All processing of the waveforms was identical to the processing in Chapter 3 (Methods Section 3.4). We computed the grand averaged signal across all subjects for the following conditions: Good Match, Good Mismatch, Bad Match and Bad Mismatch. Again, we only present data for the central presentation trials in this chapter. We also chose appropriate time-windows for analysis based on the mix of components observed in our study. The N400 analysis (350-500 ms) was added to the results of this study.

4.3 Results

The timing, polarity and scalp distributions of the effects in our experiment are consistent with N300 component (McPherson & Holcomb, 1999; Schendan and Kutas, 2002; 2003; 2007). We report the ERP amplitude in the N300 and N400 time-windows for good and bad exemplars under the two cueing conditions: Match and Mismatch.

N300: A sample set of ERP waveforms at eight channel locations are plotted in Figure 4.2.A. We replicate the N300 effect of the previous experiment for the good and bad exemplars, with a frontally distributed negativity for the bad exemplars that is greater than that for good exemplars. Because the N300 effect is frontal, we compute the mean for the four conditions, taking the average across 11 frontal channels (Figure 4.2.B, Table 4.1). This again shows that we replicate the difference between good and bad exemplars in the N300 amplitude for the match condition over frontal sites. In the mismatch condition, we see a greater negativity for the good exemplars, as compared to the good exemplars in the match condition, consistent with the idea that the N300 is indexing a match of the incoming stimulus to the template activated by the context. Moreover, there was no significant difference between the good and bad exemplars when a mismatch

occurs. This is seen in the individual waveforms across frontal electrode sites (Figure 4.2A) as well as in the mean amplitude across all frontal sites (Figure 4.2B, Table 4.1).

Are the stimuli under conditions of representativeness and cueing processed by two separate neural sources and hence indexed by two separate ERP components or by common neural sources and indexed by a single component? There is a clear interaction of representativeness and cueing in the 11 frontal electrodes tests but this does not lead to an inference of separate neural generators. If there are any differences in scalp distribution of the ERP waveforms for these two conditions (representativeness and cueing) we can infer that the neural generators could be potentially different. To compare the scalp distribution of the main effects of representativeness and cueing, we computed source voltage distributions of difference waves for the two main effects: representativeness (Bad – Good) and cueing (Mismatch – Match) (Figure 4.4). In the N300 time-window, the two main effects are qualitatively similar, with both effects frontally distributed. From this we infer that the representativeness of stimuli and their match/mismatch to verbal cues are being processed by common neural sources in the N300 time-window. Quantitatively, the main effect of representativeness is larger as compared to the cueing effect in the N300 time-window.

Table 4.1. The grand average mean values, in the N300 time-window (250-350 ms), shown for 11 frontal electrode sites (see Figure 4.2B), along with t-test and Bayes factor values. There is strong evidence (large Bayes factor) for greater negativity of the N300 for bad exemplars as compared to good exemplars when the cue matches the stimulus. When there is a mismatch between the cue and the stimulus there is no evidence (small Bayes factor) for the difference

Table 4.1 (contd.). between good and exemplars in the N300 time-window. The t-test and Bayes factor calculations were performed as $mean(Bad - Good) \neq 0$.

Condition	Cue	N	Mean (micro Volts)	Standard Deviation (micro Volts)	t(19)	p	Bayes Factor
Bad	Match	20	-7.1	0.9	-7.04	1.07E-06	16914.8
Good	Match	20	-5.1	1.1			
Bad	Mismatch	20	-6.4	1.6	-0.85	0.406	0.32
Good	Mismatch	20	-6.0	1.4			

N400: The greater negativity in the waveform for the mismatched good exemplar as compared to the matched good exemplar continues beyond the N300 time-window into the N400 time-window (Figure 4.2 A). In line with the interpretation of the N400 as indexing meaning and semantic expectancy, the good exemplars, when mismatched to the cue, show a greater negativity in the N400 time-window (350-500 ms) at frontal electrode sites, as compared to good exemplars that match the cue (Figure 4.3 and Table 4.2). The bad exemplars, in the match and mismatched condition, also show a greater negativity as compared to the good exemplars in the match condition in the N400 time-window. As the N400 distribution is fronto-central for colored pictures, we tested the frontal sites for an interaction of Good/Bad x Match/Mismatch using an ANOVA. The main effect (see Table 4.2 for mean values) of Good versus Bad is not significant: $F(1,19) = 1.05$, $p = 0.32$, epsilon (Greenhouse-Geisser) = 1). The main effect of cueing (Match versus Mismatch) is significant: $F(1,19) = 6.09$, $p = 0.023$, epsilon (Greenhouse-Geisser) = 1). The interaction of Good/Bad x Match/Mismatch is also significant and survives correction for

multiple comparisions: $F(1,19) = 14.5$, $p = 0.0012$, epsilon (Greenhouse-Geisser) = 1). We also performed pair-wise comparisons for different conditions (see Table 4.2.B). The conditions (Good Match – Good Mismatch), (Good Match – Bad Match) and (Good Match - Bad Mismatch) were significant.

Similar to the computations of voltage scalp distributions for the N300 (Figure 4.4), we computed the scalp distributions of the effect of representativeness (Bad – Good) and cueing (Mismatch – Match) in the N400 time-window. In this time-window both effects are centro-parietally distributed with a slight left laterality. This suggest that these factors are contributing to a common ERP component, the N400.

Table 4.2.A. The grand average mean values, in the N400 time-window (350-500 ms), shown for 11 frontal electrode sites (see Figure 4.3). **B.** Pair-wise comparisons for different conditions in the N400 time-window.

A

Condition	Cue	N	Mean (micro Volts)	Standard Deviation (micro Volts)
Good	Match	20	-3.5	5.3
Bad	Match	20	-5.2	5.7
Bad	Mismatch	20	-4.7	6.7
Good	Mismatch	20	-5.7	6.7

Table 4.2 B (contd.).

Test	N	t(19)	p	Bayes Factor
Good Match – Good Mismatch	20	3.82	0.001162	32.1
Good Match – Bad Match	20	5.06	6.966e-05	384.2
Good Match - Bad Mismatch	20	1.81	0.01916	2.995
Bad Match – Bad Mismatch	20	1.45	0.1647	0.57
Bad Mismatch-Good Mismatch	20	1.81	0.08692	0.91

4.4 Discussion

Our goal in this study was to determine if semantics, via a verbal cue, can dynamically modulate visual processing. In our previous experiment (Chapter 3), participants only knew that one of six categories of good and bad exemplars of images will be viewed at each trial. We postulated that a template (a term we use to describe learned statistical regularities for a category) matching process is at work and that good exemplars were facilitated because they better matched a template. In the current work, we set up expectations for a particular category on a particular trial using a verbal cue, with the aim of pre-activating a particular template, and asked if this can modulate the good versus bad effect. The word cues have no perceptual similarity to the pictures, and, hence, any template that is pre-activated by the word cue must be via the semantics of the word.

Our results provide evidence that semantics, via precuing, does indeed modulate visual processing. When the stimuli, both good and bad exemplars, match the expected template we see a difference between the good and bad exemplars in the N300 time-window, replicating the results from the previous experiment (Chapter 3) and prior work (Pietrowsky et al., 1996; Schendan and Kutas, 2002; Vo and Wolfe, 2013). When the stimuli do not match the expected

template, even when the exemplars are good exemplars (e.g. cue the word "Forest" and present an image of a good exemplar beach), we see a greater negativity for the good exemplars in the N300 time-window as compared to good exemplars in the match condition. That is, good exemplars are processed similar to bad exemplars, when they are unexpected, due to mismatch between the expected template and the incoming stimulus. Our results show that the good versus bad N300 difference is eliminated when there is a mismatch between the semantic cue and the image seen. Thus, the good versus bad effect is not driven by static templates activated only by the incoming stimulus, but by templates activated in the context of the experiment, in this case via a semantic cue. In other words, the. good versus bad effect can be modulated with expectation -- an expectation initiated dynamically via a semantic cue describing the expected category. We note that the word cues have no perceptual similarity to the stimulus and hence the template must be semantically activated. Therefore, we can initiate category templates, via cues in a different modality (language), before viewing a stimulus and this template interacts with the incoming stimulus and change our processing of the stimulus within a time-window of 250-350 ms after the stimulus onset. This provides strong evidence for the interaction of semantics and visual processing.

This also aligns well with proposed models of the N300 being a time-window for higher-order visual processing with object selection and template matching (Schendan and Kutas, 2002; 2003; 2007). With semantic precuing, subjects had sufficient time to process the cue and instantiate a template for the expected category. This can possibly be initiated by tuning of multiple brain regions to the expected stimulus (Çukur, Nishimoto, Huth, & Gallant, 2013). All these results

taken together provide strong evidence for the interaction of semantics with the visual processing of natural scenes.

The pattern of results for good and bad exemplars in the N400 time window is similar to what is seen in the N400 language literature (see review Kutas and Federmeier, 2011) for within- and outside of- category exemplars and we draw an analogy to this literature for our good and bad exemplars. The N400 indexes the structure of semantic memory (Federmeier and Kutas, 1999) with words that exactly match the expected word (e.g. in the context of sweeteners the word "*sugar*") showing the lowest N400 amplitude, while words belonging to a category outside of the expected category (e.g. *"dog"*) showing the largest N400 amplitude. Words that are not the expected word but related to the expected word (e.g. "*honey*") show an intermediate N400 amplitude. Our results show that the good exemplars that match the cue show the lowest N400 amplitude, indicating a match to the expected context. The other the conditions (good mismatch, bad mismatch and bad match) have larger N400 amplitudes, significantly different from the N400 amplitude for the good match exemplars. This indicates that the bad exemplars in the match condition and good and bad exemplars in the mismatch condition may be considered as out of category to the cue. Numerically, the Bad Match exemplars elicit N400 amplitudes that are in between those of Good Match and Good Mismatch, reminiscent of the pattern of effects for related but less expected words like "honey." However, the differences between Bad Match, Bad Mismatch and Good Mismatch are not significant and hence it is unclear whether the intermediate N400 amplitudes for exemplars related to the cue but not an exact match replicate across words and pictures. We also note that data on semantic attributes of natural scenes is currently lacking and so it is unclear how related the Bad scenes are to the Good scenes. Data

will need to be collected in the future to make a more robust model of semantic neighborhoods for multiple categories of natural scenes that will help probe the gradient of the N400 amplitude as related to expected and unexpected natural scenes.

Given the above interpretations of the N300 and the N400, one additional inference that we can make from our data and prior work is that the N300 and N400 index different things. Previous work on the N300, suggests it indexes perceptual regularities in the stimulus (Vo and Wolfe, 2013; Schendan and Kutas, 2002; 2003; 2007; McPherson & Holcomb, 1999; Pietrowsky et al., 1996), whereas the N400 indexes meaning in the stimulus (Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008). Here, in the N300 time-window, the N300 amplitudes for bad exemplars, either in the matched or mismatched condition, are overall more negative than those for the good exemplars (Figure 4.2 and difference waves in Figure 4.3). This pattern implies that N300 amplitude, despite miscuing, trends towards indexing perceptual regularities (see also Mcpherson & Holcomb, 1999); that is, the N300 shows a facilitation to good exemplars (lower N300 amplitude) even in the mismatch condition, indicating that it indexes representativeness to a well learned category, even when the stimulus does not fit the current semantic context. If the N300 amplitude indexed meaning, the amplitude would have been much larger for the good mismatched exemplars as they do not fit the semantic context of the cue. In contrast, in the N400, the good exemplars in the match condition, have the least negative amplitude and the remaining conditions (good mismatch, bad mismatch and bad match) have larger N400 amplitudes. This pattern indicates that the N400 clearly indexes the processing of the meaning of the stimulus in the context of the verbal cue, rather than the perceptual form. The good exemplars in the match and mismatch case are equally representative of their category in the two

cases. If the N400 indexed representativeness, the amplitude of the N400 would have been similar in the match and mismatch conditions. The difference in the match and mismatch conditions is the meaning of the good exemplar with respect to the semantic context provided by the cue. In the mismatch condition, the meaning of the good exemplar is completely out of context to wth respect to the semantics of the verbal cue while in the match condition it fits the semantic context. The differences between the N300 and N400 are also seen in the scalp distributions of the effects, with the N300 being frontal and the N400 for natural scenes being centro-parietal with a slight left laterality.

In conclusion, we have shown that semantics, via a verbal precue, can modulate the processing of good and bad exemplars of natural scenes in the later stages of perceptual processing, in a 250-350 ms time-window. This modulation occurs despite the written word having no perceptual resemblance to the stimulus. It is the meaning, or semantics, of the word that is modulating visual processing as the N300 time-window is considered a perceptual selection, matching or pruning stage. This, we believe, is strong evidence for the influence of semantics on visual processing on natural scenes.

4.5 Figures

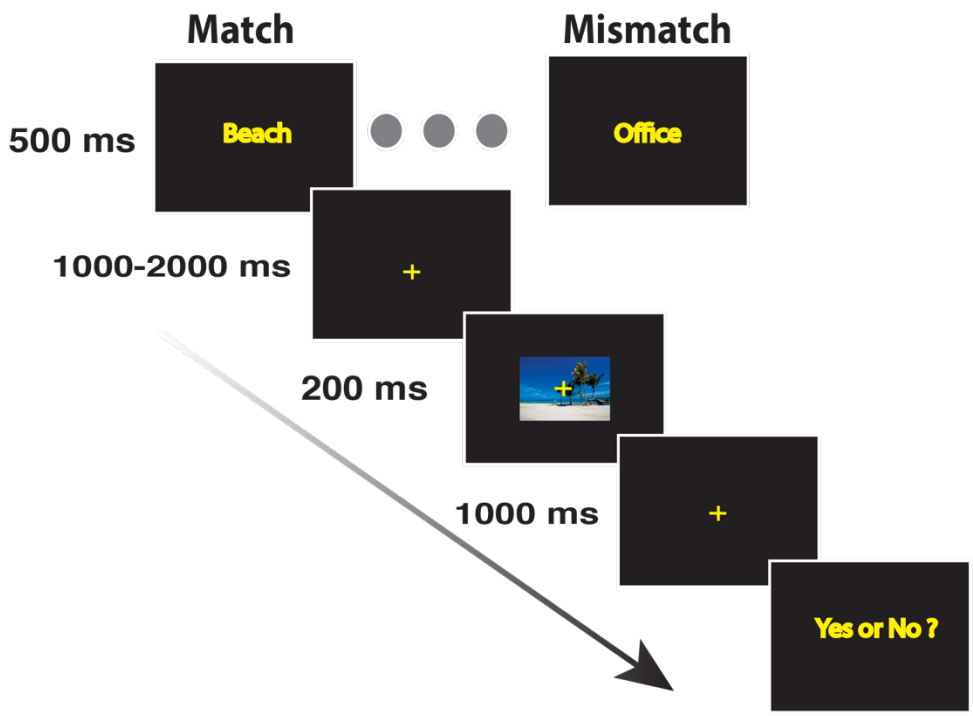
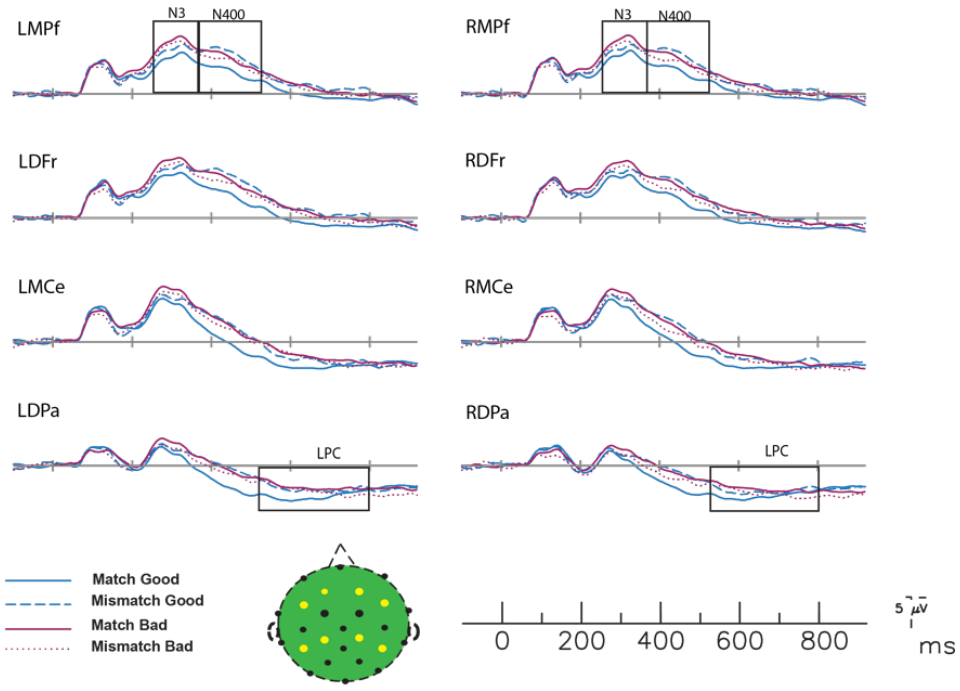


Figure 4.1. Schematic of one trial in the experiment. At the start of each trial a word cue (e.g. "Beach") from one of six categories (beaches, city streets, forests, highways, mountains, and offices) is shown. This is followed by a screen with a fixation cross is shown for time-period randomized between 1000-2000 ms. A good or bad exemplar image from one of the six categories is presented at one of the following locations: the center, the right visual field or the left visual field, followed by a fixation cross. The subjects then make a delayed response, with a button press, to the question "Yes or No?" ("Yes" if the image matches the cue and "No" otherwise) and the next trial begins. On 25% of the trials, there is a mismatch between the word cue and the image category.

Figure 4.2

A


LMPf
N3
N400
RMPf
N3
N400
LDFr
RDFr
LMCe
RMCe
LDPa
LPC
RDPa
LPC
Match Good
Mismatch Good
Match Bad
Mismatch Bad
0
200
400
600
800
ms

B

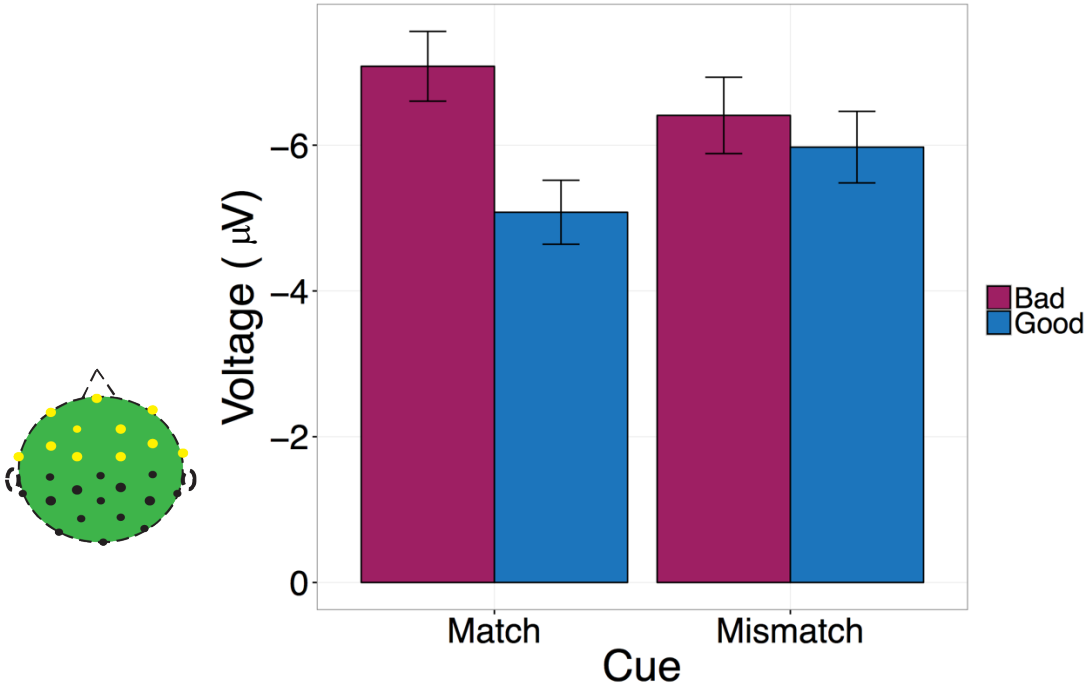

Voltage (μV)
−6
−4
−2
0
Match
Mismatch
Cue
Bad
Good

Figure 4.2 A (contd.): The grand average ERP waveforms for the good-match (solid-blue), bad-match (solid-maroon), good-mismatch (dotted-blue), bad-mismatch (dotted-maroon) conditions shown at 8 electrode sites. The channel locations are marked in yellow on the schematic of the scalp. The good/bad waveforms for the match conditions differ in the N300 time-window (250-350 ms), with greater negativity for bad exemplars as compared to good exemplars, over frontal sites. In the mismatch condition, the differences between good and bad exemplars in the N300 time-window are reduced. We also see a distinct pattern in the N400 time-window with the matched good exemplars facilitated in processing while the matched bad exemplars and the mismatched good and bad exemplars showing greater negativity than the matched good exemplar. **B**: The mean of the ERP amplitude over 11 frontal electrode sites (N = 20). The plotted error bars are within-subject confidence intervals. This shows that the difference between good and bad exemplars in the N300 time-window replicates results from the previous experiment (Figure 3.2) in the match condition. There is no evidence for differences between good and bad exemplars in the mismatch case (Bayes factor = 0.32).

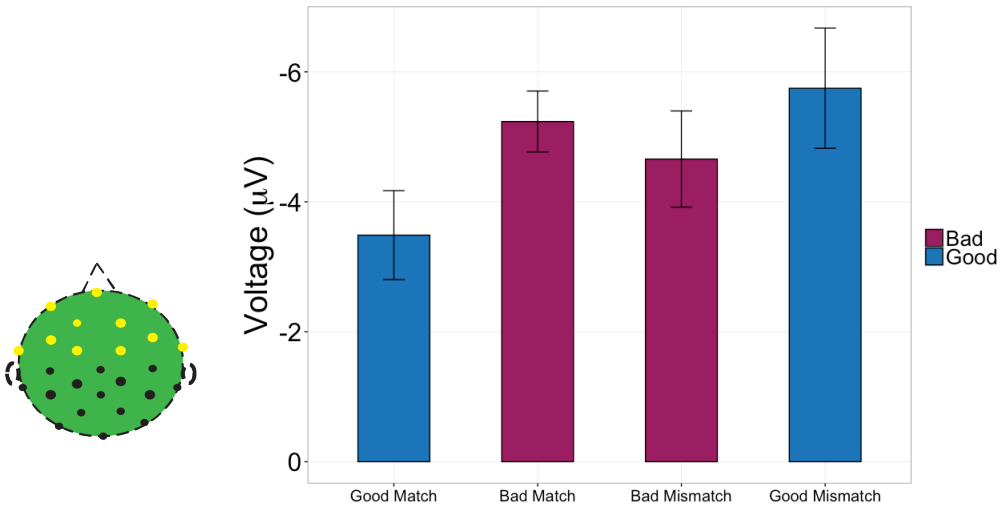


Figure 4.3: The mean of the N400 ERP amplitude over 11 frontal electrode sites (N = 20). The plotted error bars are within-subject confidence intervals. There is a significant difference between good exemplars in the match and mismatch condition, with the greatest negativity for the good mismatched exemplars. The bad exemplars show an intermediate N400 amplitude, between the good match and the good mismatch amplitudes.

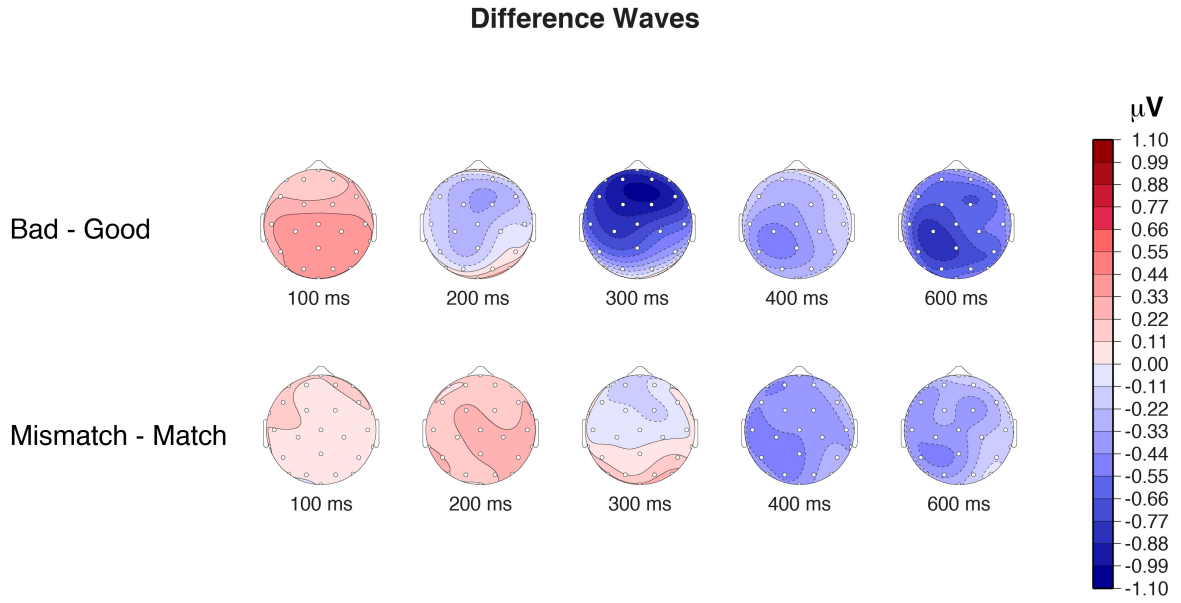


Figure 4.4: Topographic plots of the difference waves for the two main effects of representativeness (Bad – Good) and cueing (Mismatch –Match). In the N300 time-window the two main effects are qualitatively similar, with both main effects showing a frontal distribution. The N300 time-window also shows a quantitatively larger effect for the representativeness (Bad – Good) than for the cueing (Mismatch – Match). In the N400 time-window, both effects are centro-parietally distributed with a slight left laterality.

Chapter 5

Conclusion

One of the long-standing questions in cognitive science is: Does the semantics of a stimulus in one modality (vision, hearing, taste, touch and smell) influence the perception of the stimulus in another modality? In my thesis, I specifically focused on semantics as instantiated by language (written words) and sought to determine if this influenced the visual processing of natural scenes. There is no perceptual similarity between written words and pictures of natural scenes. Thus, any influence of reading the word on visual processing of images must necessarily be mediated by the meaning, or semantics, of the word and scene. I have thus used words to instantiate semantics and have shown that this instantiation is cross-modal and does influence visual processing. I summarize the results of this dissertation below.

As a first step in showing that semantics may influence visual processing, I showed that there are multiple brain regions that show similar patterns of activity between pictures and words. This extends prior work that examined brain regions that showed an overlap of areas that are activated by pictures and words. Overlap only implies common brain regions across pictures and words, but indicates nothing about the underlying patterns of activity for each modality. Our results showing similar activity patterns across modalities imply that the underlying neural code is similar for pictures and words describing natural scenes. This is stronger evidence for the existence of a common representation across pictures and words. These regions that show similar representations across modalities have also been implicated in studies using words and objects, with some commonalities and differences. There are many common regions that we see in our studies and those in the other studies using simple objects. The angular gyrus, IPL, Precuneus

and lateral posterior temporal regions are common across studies, indicating a core semantic network. The differences that we see with our study as compared to other studies are in higher visual areas: we see many brain regions that are specific to scene processing such as the anterior PPA and the RSC, while other studies (Fairhall and Caramazza, 2013) using objects do not show these regions as sharing similar representations. This implies that these regions show similar patterns of activation, across pictures and words, based on the type of stimulus (scenes as opposed to objects). We can thus speculate that these regions preferentially process stimuli (e.g. PPA for natural scenes) and hence are perceptual regions that are activated by words that trigger the semantics relevant to these regions (e.g. words evoking scenes for the anterior PPA). So, in addition to a core semantic network, we have a distributed set of regions in higher visual areas that show similar patterns of activity across pictures and words and also show preference to the type of stimulus. It is these brain regions that plausibly could act as processing sites where semantics and visual processing interact.

To examine if there could be a dynamic interaction between semantics and visual processing, we used representativeness of a scene as a proxy for prior knowledge. The properties that make a scene representative (good) for a category have been well learnt over time. Although prior work showed that representativeness mattered in categorization and detection tasks (Caddigan et al., 2010; 2017), the information was lacking on whether this was due to low-level visual features (e.g. color, contrast, luminance), high-level visual features (e.g. spatial structure) or due to the semantic information pertaining to the representativeness of the scene. Using ERPs we determined that the brain is sensitive to representativeness of scenes in a time-window (250-350 ms) indexed by the N300 component. This time-window is later than the time scale for low-level

visual processing and consistent with higher level visual processing, at the perceptual selection and matching stage for global visual properties (Schendan and Kutas, 2002; 2003; 2007). Good scenes are facilitated in processing as compared to bad scenes (Chapter 3). In comparison to prior work with objects, we can infer that a match to a template has been made at this stage for good and bad exemplars in the N300 time-window. Based on our own results and prior work on the N300 (Schendan and Kutas, 2002), we linked the N300 to a stage of processing that is sensitive to learned statistical regularities, encompassing global structure (Schendan and Kutas, 2002), canonical viewpoints (Schendan and Kutas, 2003), probableness (Vo and Wolfe, 2013), and representativeness of scenes (in this work, Chapters 3 and 4). We argue that good exemplars are facilitated because they match statistical regularities (or templates) that we have learned over time. This result still does not tell us if these templates, that come from our prior knowledge, can interact with visual processing. Also, we do not know if these templates are instantiated only by perceptual stimuli or if can they be activated cross-modally, for example by words.

To better understand the dynamics with which semantic information might influence visual processing, we used words to try to pre-activate templates. The words provided cues to the category of the incoming stimulus. When the incoming stimuli matched expectations, there were differences in the ERP waveform between good and bad exemplars in the N300 time-window (250-350 ms). When there was a mismatch between expectation and the incoming stimuli, there was no significant difference in the ERP waveform between the good and bad exemplars in the N300 time-window. From these results, we inferred that the word cues help instantiate category templates and that good exemplars were facilitated in processing because they better matched the templates for their categories. Under conditions of mismatch, when good scenes, which are

representative of a different category than the cue, are processed, they do not match the active templates instantiated by the cue. This results in even these good scenes being processed as bad scenes due to the mismatch with the current template. The mechanism of instantiating category templates via written words is necessarily mediated through semantics, as the written words have no perceptual similarity to the images. Thus, semantic information via words has dynamically influenced visual processing of natural scene images.

The dynamics of the interaction of semantics with visual processing, differ is some aspects between when there is a word pre-cue as compared to when there is no pre-cue during trials. Although in both cases the stimulus is matched to learned regularities, there are differences in the activation of these regularities. When each trial begins with a verbal cue, participants can pre-activate templates, before the stimulus is shown. The visual processing that follows interacts with this pre-activated template. When there is no verbal cue at the beginning of the trial, the stimulus is still matched to known learned regularities, but how are these regularities activated? One possible mechanism is via the extraction of scene gist. Humans are very efficient at extracting the gist of scenes (Walther et al., 2009), even in low resolution conditions when individual objects in the scene cannot be identified (Oliva and Schyns, 1995). This gist extraction occurs in a time-window after about 150 ms (Thorpe et al., 1996). Once the gist is extracted, a template corresponding to the gist can be activated and the incoming stimulus can be matched to this template, which we argue occurs in the N300 time-window. Thus, in the two experiments, one with a pre-cue and one without a pre-cue, the incoming stimulus can be matched to activated templates and dynamically processed.

One additional factor that we must consider from these experiments is the differing methodologies used: fMRI and ERP. While fMRI has high spatial resolution and low temporal resolution, ERPs have high temporal resolution and low spatial resolution. The results from the fMRI thus show us information at long time scales. Hence, the spatial regions that we see in our fMRI study include processing at long time-scales, i.e. both feedback and recurrent processing. The ERP results give us information at fast time-scales of processing. Given that we see the interaction of semantics and visual processing from the ERP waveforms in the N300 time-window, and the fMRI results indicated cross-decoding in higher visual areas, we believe that the higher visual areas are a plausible site for the interaction of semantics and visual processing. Other studies (Hasson et al., 2015) have implicated high-level visual processing regions as processing information at longer time-scales and have also implicated them in cross-modal processing (Chen et al., 2016).

We have thus provided evidence for the dynamic influence of semantics on visual processing of natural scenes and also showed brain regions that represent information similarly across pictures and words. Semantic cueing, via words, can set up expectancy and modulate the processing of incoming pictures. This interactive processing can occur in brain regions processing semantics linked to different visual processing regions or it can occur in local regions, where we showed that words and pictures had similar patterns of representation. Indeed, the distinction between semantic information and visual information may well be artificial in some brain regions, as their representations can be interchanged. More work would be needed to delineate the nature of processing in regions that show similar patterns of activity across pictures and words. There is no doubt though, that in some situations, semantics does influence visual processing.

References

Baayen, R. H., Piepenbrock, R., & Rijn, van H. (1993). The CELEX lexical data base on CD-ROM.

Bar, M., & Aminoff, E. (2003). Cortical Analysis of Visual Context. *Neuron*, *38*, 347–358.

Bar, M., K. S. Kassam, A. S. Ghuman, J. Boshyan, A. M. Schmid, A. M. Dale, M. S. Hämäläinen, K. Marinkovic, D. L. Schacter, and B. R. Rosen. "Top-down Facilitation of Visual Recognition." Proceedings of the National Academy of Sciences of the United States of America 103, no. 2 (2006): 449–454.

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2013). Differential connectivity within the Parahippocampal Place Area. NeuroImage, 75, 228–237. https://doi.org/10.1016/j.neuroimage.2013.02.073

Baldassano, C., Fei-Fei, L., & Beck, D. M. (2016). Pinpointing the peripheral bias in neural scene-processing networks during natural viewing. *Journal of Vision*, *16*(2), 9. https://doi.org/10.1167/16.2.9

Baldassano, C., Esteva, A., Fei-Fei, L., & Beck, D. M. (2016). Two Distinct Scene-Processing Networks Connecting Vision and Memory. *eNeuro*, *3*(5). https://doi.org/10.1523/ENEURO.0178-16.2016

Barsalou, L. W., Kyle Simmons, W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in Cognitive Sciences, 7(2), 84–91. http://doi.org/10.1016/S1364-6613(02)00029-3

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct Brain Systems for Processing Concrete and Abstract Concepts. Journal of Cognitive Neuroscience, 17(6), 905–917. http://doi.org/10.1162/0898929054021102

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cerebral Cortex, 19(12), 2767–2796. doi:10.1093/cercor/bhp055

Binney, R. J., Embleton, K. V., Jefferies, E., Parker, G. J. M., & Lambon Ralph, M. A. (2010). The Ventral and Inferolateral Aspects of the Anterior Temporal Lobe Are Crucial in Semantic Memory: Evidence from a Novel Direct Comparison of Distortion-Corrected fMRI, rTMS, and Semantic Dementia. *Cerebral Cortex*, *20*(11), 2728–2738. https://doi.org/10.1093/cercor/bhq019

Borghesani, V., Pedregosa, F., Buiatti, M., Amadon, A., Eger, E., & Piazza, M. (2016). Word meaning in the ventral visual path: a perceptual to conceptual gradient of semantic coding. *NeuroImage*, *143*, 128–140. https://doi.org/10.1016/j.neuroimage.2016.08.068

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10(4), 433–436.

Bray, S., Arnold, A. E. G. F., Levy, R. M., & Iaria, G. (2015). Spatial and temporal functional connectivity changes between resting and attentive states: Connectivity Changes Between Rest and Attention. *Human Brain Mapping*, *36*(2), 549–565. https://doi.org/10.1002/hbm.22646

Breitmeyer, B. G., & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review, and update. *Perception & Psychophysics*, *62*(8), 1572–1595. https://doi.org/10.3758/BF03212157

Bruffaerts, R., Dupont, P., Peeters, R., Deyne, S. D., Storms, G., & Vandenberghe, R. (2013). Similarity of fMRI Activity Patterns in Left Perirhinal Cortex Reflects Semantic Similarity between Words. The Journal of Neuroscience, 33(47), 18597–18607. http://doi.org/10.1523/JNEUROSCI.1548-13.2013

Caddigan, E., Walther, D. B., Fei-Fei, L., & Beck, D. M. (2010). Perceptual differences between natural scene categories. OPAM 2010 18th Annual Meeting, Visual Cognition,18(10), 1498-1502.

Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of Vision*, *17*(1), 21. https://doi.org/10.1167/17.1.21

Caramazza, A., Hillis, A. E., Rapp, B. C., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology*, *7*(3), 161–189. https://doi.org/10.1080/02643299008253441

Carmel, D., & Bentin, S. (2002). Domain specificity versus expertise: factors influencing distinct processing of faces. *Cognition*, *83*(1), 1–29. https://doi.org/10.1016/S0010-0277(01)00162-7

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol., 2(3), 27:1–27:27. http://doi.org/10.1145/1961189.1961199

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nature Neuroscience, 2(10), 913–919.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115-125.

Clarke, A., & Tyler, L. K. (2014). Object-Specific Semantic Coding in Human Perirhinal Cortex. The Journal of Neuroscience, 34(14), 4766–4775. http://doi.org/10.1523/JNEUROSCI.2828-13.2014

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., 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), 2608–2618. http://doi.org/10.1523/JNEUROSCI.5547-11.2012

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. Computers and Biomedical Research, 29(3), 162–173. http://doi.org/10.1006/cbmr.1996.0014

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201. https://doi.org/10.1037/0096-3445.132.2.163

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770. https://doi.org/10.1038/nn.3381

Dale, A. M. (1994). Source localization and spatial discriminant analysis of event-related potentials: linear approaches (brain cortical surface). Dissertation Abstracts International.

David, O., Harrison, L., & Friston, K. J. (2005). Modelling event-related responses in the brain. *NeuroImage*, *25*(3), 756–770. https://doi.org/10.1016/j.neuroimage.2004.12.030

Delis, D. C., Robertson, L. C., & Efron, R. (1986). Hemispheric specialization of memory for visual hierarchical stimuli. *Neuropsychologia*, *24*(2), 205–214.

D'Esposito, M., Detre, J. A., Aguirre, G. K., Stallcup, M., Alsop, D. C., Tippet, L. J., & Farah, M. J. (1997). A functional MRI study of mental image generation. Neuropsychologia, 35(5), 725–730. doi:10.1016/S0028-3932(96)00121-2

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. The Journal of Neuroscience, 33(48), 18906–18916. doi:10.1523/JNEUROSCI.3809-13.2013

Dobbins, I. G., Kroll, N. E., Tulving, E., Knight, R. T., & Gazzaniga, M. S. (1998). Unilateral medial temporal lobe memory impairment: type deficit, function deficit, or both? *Neuropsychologia*, *36*(2), 115–127.

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R. J., & Rees, G. (2008). fMRI Activity Patterns in Human LOC Carry Information about Object Exemplars within Category. *Journal of Cognitive Neuroscience*, *20*(2), 356–370. https://doi.org/10.1162/jocn.2008.20019

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

Evans, K. M., & Federmeier, K. D. (2007). The memory that's right and the memory that's left: Event-related potentials reveal hemispheric asymmetries in the encoding and retention of verbal information. Neuropsychologia, 45(8), 1777–1790. http://doi.org/10.1016/j.neuropsychologia.2006.12.014

Fairhall, S. L., & Caramazza, A. (2013). Brain Regions That Represent Amodal Conceptual Knowledge. Journal of Neuroscience, 33(25), 10552–10558. doi:10.1523/JNEUROSCI.0051-13.2013

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., & Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 202–224. https://doi.org/10.1037//0278-7393.27.1.202

Federmeier K. D., Kutas M., Dickson D. S. (2015). A common neural progression to meaning in about a third of a second. In: Hickok GS, Small SL, editors. Neurobiology of Language. (pp. 557-568). Holland: Elsevier.

Finnigan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). ERP "old/new"effects: memory strength and decisional factor (s). *Neuropsychologia*, *40*(13), 2288–2304.

Firestone, C., & Scholl, B. J. (2014). "Top-Down" Effects Where None Should Be Found: The El Greco Fallacy in Perception Research. *Psychological Science*, *25*(1), 38–46. https://doi.org/10.1177/0956797613485092

Friedman, L., Glover, G. H., Krenz, D., & Magnotta, V. (2006). Reducing inter-scanner variability of activation in a multicenter fMRI study: Role of smoothness equalization. *NeuroImage*, *32*(4), 1656–1668. https://doi.org/10.1016/j.neuroimage.2006.03.062

Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. Cognitive Brain Research, 20(2), 226–241. http://doi.org/10.1016/j.cogbrainres.2004.02.012

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363. https://doi.org/10.1038/nrn3476

Glaser, W. R. (1992). Picture naming. Cognition, 42(1), 61–105.

Gonzalez, C. M. G., Clark, V. P., Fan, S., Luck, S. J., & Hillyard, S. A. (1994). Sources of attention-sensitive visual event-related potentials. *Brain Topography*, *7*(1), 41–51.

Greene, M.R., Botros, A.P., Beck, D.M , Fei-Fei, L. (2015) What You See is What You Expect: Rapid Scene Understanding Requires Prior Experience. Attention, Perception, & Psychophysics, 77, (4), 1239-1251.

Greene, M., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cognitive Psychology, 58(2), 137–176. doi:10.1016/j.cogpsych.2008.06.001

Grill-Spector, K., & Kanwisher, N. (2005). Visual Recognition: As Soon as You Know It Is There, You Know What It Is. *Psychological Science*, *16*(2), 152–160.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., &

Hasson U, Chen J, Honey CJ (2015). Hierarchical process memory: memory as an integral component of information processing. Trends in Cognitive Sciences 19:304-313.

Holcomb, P. J., & Mcpherson, W. B. (1994). Event-Related Brain Potentials Reflect Semantic Priming in an Object Decision Task. *Brain and Cognition*, *24*(2), 259–276. https://doi.org/10.1006/brcg.1994.1014

Hsieh, P.-J., Vul, E., & Kanwisher, N. (2010). Recognition Alters the Spatial Pattern of fMRI Activation in Early Retinotopic Cortex. Journal of Neurophysiology, 103(3), 1501–1507. http://doi.org/10.1152/jn.00812.2009

Jeong, S. K. & Xu, Y. (2016). Behaviorally relevant abstract object identity representation in the human parietal cortex. Journal of Neuroscience, 36, 1607-1619. (pdf)

Kan, I. P., Barsalou, L. W., Olseth Solomon, K., Minor, J. K., & Thompson-Schill, S. L. (2003). Role of Mental Imagery in a Property Verification Task: Fmri Evidence for Perceptual Representations of Conceptual Knowledge. Cognitive Neuropsychology, 20(3-6), 525–540. http://doi.org/10.1080/02643290244000257

Kaplan, J. T., Man, K., & Greening, S. G. (2015). Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. Frontiers in Human Neuroscience, 9. http://doi.org/10.3389/fnhum.2015.00151

Kherif, F., Josse, G., & Price, C. J. (2010). Automatic Top-Down Processing Explains Common Left Occipito-Temporal Responses to Visual Words and Objects. Cerebral Cortex, 21, 103–114. http://doi.org/10.1093/cercor/bhq063

Kleiner, M., Brainard, D., Pelli, D. G., Ingling, A., Murray, R., & Broussard, C. M. (2007). What's new in Psychtoolbox-3. Perception, 36(14), 1. Perception, 36(14), 1.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent Processing in V1/V2 Contributes to Categorization of Natural Scenes. The Journal of Neuroscience, 31(7), 2488–2492. doi:10.1523/JNEUROSCI.3074-10.2011

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*(4), 217–230. https://doi.org/10.1038/nrn3008

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

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

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kumar, M., Federmeier, K. D., Fei-Fei, L., & Beck, D. M. (2017). Evidence For Similar Patterns of Neural Activity Elicited by Picture- and Word-based Representations of Natural Scenes. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.03.037

Kutas, M. (1993). In The Company of Other Words: Electrophysiological Evidence For Single-word and Sentence Context Effects. *Language and Cognitive Processes*, *8*(4), 533–572.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

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

Kveraga, K., Boshyan, J., & Bar, M. (2007). Magnocellular Projections as the Trigger of Top-Down Facilitation in Recognition. *Journal of Neuroscience*, *27*(48), 13232–13240. https://doi.org/10.1523/JNEUROSCI.3481-07.2007

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and Differentiation in Semantic Memory: Insights from Semantic Dementia. Annals of the New York Academy of Sciences, 1124(1), 61–76. http://doi.org/10.1196/annals.1440.006

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9596–9601. https://doi.org/10.1073/pnas.092277599

Liuzzi, A. G., Bruffaerts, R., Dupont, P., Adamczuk, K., Peeters, R., De Deyne, S., … Vandenberghe, R. (2015). Left perirhinal cortex codes for similarity in meaning between written words: Comparison with auditory word input. Neuropsychologia, 76, 4–16. http://doi.org/10.1016/j.neuropsychologia.2015.03.016

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513–536. https://doi.org/10.1080/13506280902937606

Luck, S. J., & Kappenman, E. S. (2011). *The Oxford Handbook of Event-Related Potential Components*. Oxford University Press.

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond Gist: Strategic and Incremental Information Accumulation for Scene Categorization. *Psychological Science*, *25*(5), 1087–1097. https://doi.org/10.1177/0956797614522816

Martin, A. (2007). The Representation of Object Concepts in the Brain. Annual Review of Psychology, 58(1), 25–45. http://doi.org/10.1146/annurev.psych.57.102904.190143

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. Current Opinion in Neurobiology, 11(2), 194–201.

McRae, K., Cree, G. S., Seidenberg, M. S., & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

Mesulam, M.-M., Wieneke, C., Hurley, R., Rademaker, A., Thompson, C. K., Weintraub, S., & Rogalski, E. J. (2013). Words and objects at the tip of the left temporal lobe in primary progressive aphasia. Brain, 136(2), 601–618. doi:10.1093/brain/aws336

Montaldi, D., Spencer, T. J., Roberts, N., & Mayes, A. R. (2006). The neural system that mediates familiarity memory. *Hippocampus*, *16*(5), 504–520. https://doi.org/10.1002/hipo.20178

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005)." *Reason*, 4.2 (2008):61.

Mukherjee, S., Golland, P., & Panchenko, D. (2003). Permutation tests for classification. Retrieved from http://18.7.29.232/handle/1721.1/6723

Münte, T. F., Urbach, T. P., Düzel, E., & Kutas, M. (2000). Event-related brain potentials in the study of human cognition and neuropsychology. *Handbook of Neuropsychology*, *1*, 139–235.

Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience*, *4*(1), 15–22.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Comput Biol*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stiimulus-specific brain regions. Journal of Cognitive Neuroscience, 12(6), 1013–1023.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Oliva, A. (2005). Gist of the scene. Neurobiology of Attention, 696(64), 251–258.

Oliva, A., & Schyns, P. G. (1995). Mandatory scale perception promotes flexible scene categorization. In *Proceedings of the XVII Meeting of the Cognitive Science Society. Lawrence Erlbaum, Hillsdale, NJ* (pp. 159–163).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research: Visual Perception*, *155*, 23–36.

Paivio, A. (1974). Language and Knowledge of the World. *Educational Researcher*, *3*(9), 5. https://doi.org/10.2307/1174914

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience, 8(12), 976–987. doi:10.1038/nrn2277

Pelli, D. G. (1998). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437–442.

Pietrowsky, R., Kuhmann, W., Krug, R., Molle, M., Fehm, H. L., & Born, J. (1996). Event-related brain potentials during identification of tachistoscopically presented pictures. *Brain and Cognition*, *32*(3), 416–428.

Price, C. J. (2000). The anatomy of language: contributions from functional neuroimaging. *Journal of Anatomy*, *197*(03), 335–359. https://doi.org/null

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience, 11(5), 351–360.

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(03), 341–365.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rosch, E., CB, M., WD, G., DM, J., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8–3.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*. https://doi.org/10.1038/nn866

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes. *Visual Cognition*, *12*, 852–877.

Schechter, I., Butler, P. D., Zemon, V. M., Revheim, N., Saperstein, A. M., Jalbrzikowski, M., … Javitt, D. C. (2005). Impairments in generation of early-stage transient visual evoked potentials to magno- and parvocellular-selective stimuli in schizophrenia. *Clinical Neurophysiology*, *116*(9), 2204–2215. https://doi.org/10.1016/j.clinph.2005.06.013

Schendan, H. E., & Kutas, M. (2002). Neurophysiological evidence for two processing times for visual object identification. *Neuropsychologia*, *40*(7), 931–945.

Schendan, H. E., & Kutas, M. (2003). Time Course of Processes and Representations Supporting Visual Object Identification and Memory. *Journal of Cognitive Neuroscience*, *15*(1), 111–135. https://doi.org/10.1162/089892903321107864

Schendan, H. E., & Kutas, M. (2007). Neurophysiological evidence for the time course of activation of global shape, part, and local contour representations during visual object categorization and memory. *Journal of Cognitive Neuroscience*, *19*(5), 734–749.

Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2006). A reevaluation of the electrophysiological correlates of expert object processing. *Journal of Cognitive Neuroscience*, *18*(9), 1453–1465.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429. https://doi.org/10.1073/pnas.0700622104

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. Neuroimage, 54, 2418–2425.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, *20*(11), 2037–2057.

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. J. (2014). Modality-Independent Decoding of Semantic Information from the Human Brain. Cerebral Cortex, 24(2), 426–434. http://doi.org/10.1093/cercor/bhs324

Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: inferring "how" from "where." Neuropsychologia, 41(3), 280–292. http://doi.org/10.1016/S0028-3932(02)00161-6.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good Exemplars of Natural Scene Categories Elicit Clearer Patterns than Bad Exemplars but Not Greater BOLD Activity. *PLoS ONE*, *8*(3), e58594. https://doi.org/10.1371/journal.pone.0058594

Vandenberghe, R., Price, C., Wise, R., Josephs, O., & Frackowiak, R. S. J. (1996). Functional anatomy of a common semantic system for words and pictures. Nature, 383(6597), 254–256. http://doi.org/10.1038/383254a0

Visser, M., Jefferies, E., & Ralph, M. L. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. Journal of Cognitive Neuroscience, 22(6), 1083–1094.

Vo, M. L.-H., & Wolfe, J. M. (2013). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, *24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. Journal of Neuroscience, 29(34), 10573–10581. doi:10.1523/JNEUROSCI.0559-09.2009

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853.

Willems, R. M., Özyürek, A., & Hagoort, P. (2008). Seeing and Hearing Meaning: ERP and fMRI Evidence of Word versus Picture Integration into a Sentence Context. *Journal of Cognitive Neuroscience*, *20*(7), 1235–1249. https://doi.org/10.1162/jocn.2008.20085

Appendix A

List of Phrases Describing Natural Scenes

List of phrases used as word stimuli. Each box contains the 4-phrase sets that appeared in a block. The sequence of phrases within a block and the block order were randomized across subjects.

Beaches

1	APPEALING	BEACH	33	PEACEFUL	BEACH
2	BEAUTIFUL	SEASIDE	34	LOVELY	SEASHORE
3	WINDY	SEASHORE	35	SERENE	BEACH
4	QUIET	BEACH	36	PICTURESQUE	SEASIDE
5	BEAUTIFUL	SEASHORE	37	SUNNY	SEASHORE
6	SUNNY	SEASIDE	38	PLEASANT	BEACH
7	BREEZY	BEACH	39	RADIANT	SEASIDE
8	GOLDEN	SEASHORE	40	PLEASING	SEASIDE
9	BREATHTAKING	SEASIDE	41	TROPICAL	SEASHORE
10	VAST	SEASIDE	42	RESPLENDENT	SEASIDE
11	BRIGHT	SEASHORE	43	SANDY	BEACH

Beaches (contd.)

12	SUNNY	BEACH	44	REFRESHING	SEASHORE
13	CLOUDY	SEASIDE	45	TROPICAL	SEASIDE
14	SCENIC	BEACH	46	RELAXING	BEACH
15	BREEZY	SEASHORE	47	PRETTY	SEASHORE
16	DELIGHTFUL	SEASIDE	48	ROMANTIC	BEACH
17	DESOLATE	BEACH	49	SANDY	SEASIDE
18	ENTICING	SEASHORE	50	SCENIC	SEASHORE
19	WINDY	BEACH	51	SERENE	SEASHORE
20	GRAND	SEASIDE	52	SOOTHING	BEACH
21	HEAVENLY	SEASHORE	53	SUBLIME	SEASHORE
22	HOT	BEACH	54	BLISSFUL	SEASHORE
23	ENTICING	BEACH	55	TRANQUIL	SEASIDE
24	HUMID	SEASIDE	56	VAST	BEACH
25	PICTURESQUE	SEASHORE	57	PICTURESQUE	BEACH
26	BREEZY	SEASIDE	58	LOVELY	SEASIDE
27	LOVELY	BEACH	59	WINDY	SEASIDE
28	WARM	SEASIDE	60	WONDERFUL	SEASHORE

Beaches (contd.)

29	LUSH	BEACH	61	SERENE	SEASIDE
30	TROPICAL	BEACH	62	ENTICING	SEASIDE
31	SCENIC	SEASIDE	62	BEAUTIFUL	BEACH
32	VAST	SEASHORE	64	SANDY	SEASHORE

Cities

1	GREEN	CITY	33	HARMONIOUS	DOWNTOWN
2	HISTORIC	CITY	34	EXPENSIVE	TOWN
3	BIG	DOWNTOWN	35	HUGE	TOWN
4	UNIFORM	TOWN	36	INDUSTRIOUS	CITY
5	BRIGHT	DOWNTOWN	37	HARMONIOUS	TOWN
6	BUSY	TOWN	38	LIVELY	TOWN
7	CENTRAL	CITY	39	HECTIC	CITY
8	PROSPEROUS	DOWNTOWN	40	GLAMOROUS	DOWNTOWN
9	CHAOTIC	DOWNTOWN	41	EXPENSIVE	CITY
10	CHARMING	TOWN	42	LIVELY	DOWNTOWN

Cities (contd.)

11	CLEAN	CITY	43	LOUD	TOWN
12	PACKED	TOWN	44	MODERN	DOWNTOWN
13	WEALTHY	CITY	45	NOISY	CITY
14	CROWDED	CITY	46	ORGANIZED	DOWNTOWN
15	VIBRANT	TOWN	47	GLAMOROUS	TOWN
16	ELECTRIC	DOWNTOWN	48	PACKED	CITY
17	DYNAMIC	TOWN	49	EXCITING	TOWN
18	POPULOUS	CITY	50	QUAINT	TOWN
19	EXCITING	DOWNTOWN	51	RESTLESS	DOWNTOWN
20	WEALTHY	TOWN	52	TALL	CITY
21	ENCHANTING	DOWNTOWN	53	ENCHANTING	TOWN
22	CROWDED	DOWNTOWN	54	PROSPEROUS	CITY
23	ENERGETIC	CITY	55	CHARMING	CITY
24	ENTICING	TOWN	56	WEALTHY	DOWNTOWN
25	PACKED	DOWNTOWN	57	LIVELY	CITY
26	ENCHANTING	CITY	58	PROSPEROUS	TOWN
27	EXCITING	CITY	59	SPARKLING	DOWNTOWN
28	FANCY	TOWN	60	CROWDED	TOWN

Cities (contd.)

29	EXPENSIVE	DOWNTOWN	61	HARMONIOUS	CITY
30	GLAMOROUS	CITY	62	VIBRANT	CITY
31	VIBRANT	DOWNTOWN	62	CHARMING	DOWNTOWN
32	GRAND	TOWN	64	CONCRETE	TOWN

Highways

1	ARTERIAL	HIGHWAY	33	MONOTONOUS	FREEWAY
2	BEAUTIFUL	FREEWAY	34	CONCRETE	INTERSTATE
3	BORING	HIGHWAY	35	MUNDANE	INTERSTATE
4	FAST	INTERSTATE	36	NECESSARY	HIGHWAY
5	BROAD	HIGHWAY	37	WIDE	INTERSTATE
6	MONOTONOUS	INTERSTATE	38	NOISY	FREEWAY
7	COMPLEX	FREEWAY	39	CONGESTED	INTERSTATE
8	CONCRETE	HIGHWAY	40	MODERN	HIGHWAY
9	WINDING	FREEWAY	41	OPEN	INTERSTATE
10	CONGESTED	FREEWAY	42	ORGANIZED	HIGHWAY

Highways (contd.)

11	SMOOTH	HIGHWAY	43	PACKED	FREEWAY
12	CROWDED	INTERSTATE	44	CONCRETE	FREEWAY
13	NOISY	INTERSTATE	45	SMOOTH	INTERSTATE
14	DIRTY	HIGHWAY	46	WINDING	HIGHWAY
15	DRAB	FREEWAY	47	EMPTY	FREEWAY
16	DREARY	INTERSTATE	48	PACKED	HIGHWAY
17	EFFICIENT	HIGHWAY	49	PERVASIVE	FREEWAY
18	EMPTY	INTERSTATE	50	PLEASURABLE	HIGHWAY
19	PANORAMIC	INTERSTATE	51	ROUGH	INTERSTATE
20	BROAD	FREEWAY	52	SLEEK	FREEWAY
21	CONGESTED	HIGHWAY	53	BROAD	INTERSTATE
22	FAST	HIGHWAY	54	SMOOTH	FREEWAY
23	FLOWING	FREEWAY	55	MONOTONOUS	HIGHWAY
24	ARTERIAL	INTERSTATE	56	SPEEDY	HIGHWAY
25	EMPTY	HIGHWAY	57	STANDARD	FREEWAY
26	FUNCTIONAL	FREEWAY	58	FAST	FREEWAY
27	PACKED	INTERSTATE	59	TEDIOUS	INTERSTATE
28	INTRUSIVE	FREEWAY	60	NOISY	HIGHWAY

Highways (contd.)

29	LINEAR	HIGHWAY	61	URBAN	HIGHWAY
30	LONELY	INTERSTATE	62	WIDE	FREEWAY
31	CLEAN	FREEWAY	62	ARTERIAL	FREEWAY
32	WIDE	HIGHWAY	64	WINDING	INTERSTATE

Mountains

1	ALPINE	PEAK	33	TOWERING	PEAK
2	BREATHTAKING	PEAK	34	ICY	MOUNTAIN
3	CHALLENGING	SUMMIT	35	LUSH	SUMMIT
4	CHILLY	MOUNTAIN	36	MAJESTIC	PEAK
5	WONDROUS	MOUNTAIN	37	LARGE	MOUNTAIN
6	MAJESTIC	SUMMIT	38	PHENOMENAL	PEAK
7	CLOUDY	PEAK	39	GLORIOUS	SUMMIT
8	COLD	SUMMIT	40	RELAXING	MOUNTAIN
9	COLOSSAL	PEAK	41	REFRESHING	MOUNTAIN
10	DANGEROUS	PEAK	42	SCENIC	MOUNTAIN
11	WILD	SUMMIT	43	ROCKY	SUMMIT

Mountains (contd.)

12	FROSTY	MOUNTAIN	44	STAGGERING	PEAK
13	LUSH	PEAK	45	ROLLING	MOUNTAIN
14	DANGEROUS	SUMMIT	46	ROUGH	PEAK
15	GIGANTIC	MOUNTAIN	47	PICTURESQUE	SUMMIT
16	ENORMOUS	MOUNTAIN	48	SENSATIONAL	PEAK
17	GIGANTIC	SUMMIT	49	SERENE	SUMMIT
18	TOWERING	SUMMIT	50	LUSH	MOUNTAIN
19	STEEP	PEAK	51	MONUMENTAL	PEAK
20	HUGE	MOUNTAIN	52	HUGE	SUMMIT
21	GIGANTIC	PEAK	53	STAGGERING	SUMMIT
22	HIGH	SUMMIT	54	TREMENDOUS	MOUNTAIN
23	STAGGERING	MOUNTAIN	55	STEEP	MOUNTAIN
24	ICY	PEAK	56	IMMENSE	PEAK
25	STEEP	SUMMIT	57	STUPENDOUS	PEAK
26	MAJESTIC	MOUNTAIN	58	DANGEROUS	MOUNTAIN
27	GLORIOUS	PEAK	59	TRANQUIL	SUMMIT
28	IMMENSE	SUMMIT	60	TOWERING	MOUNTAIN
29	GLORIOUS	MOUNTAIN	61	HUGE	PEAK

Mountains (contd.)

30	SNOWY	PEAK	62	ICY	SUMMIT
31	IMMENSE	MOUNTAIN	62	VAST	MOUNTAIN
32	CHILLY	SUMMIT	64	CHILLY	PEAK

Appendix B

Cross-Decoding Map: 400 Voxel Cluster

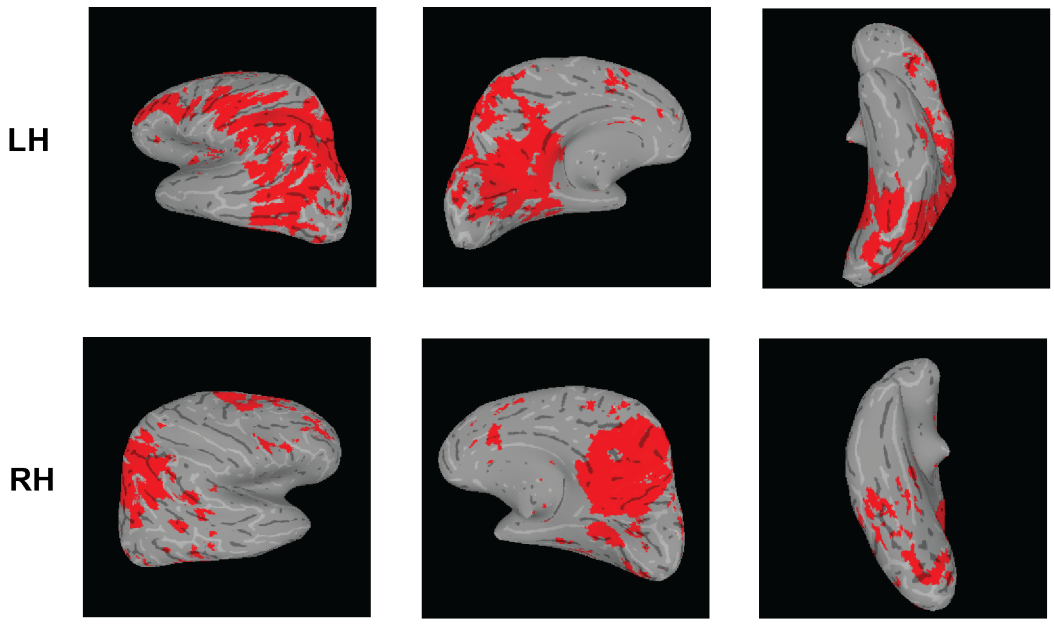


Figure B.1. Cross-decoding intersection maps created based on the permutation test cluster size of 400 clusters. A minimum threshold of 27% was determined to be significant ($p < 0.05$) for a 400 voxel cluster. We see a few more regions in the ventral-temporal areas as compared to using a cluster size of 100 voxels and a minimum threshold of 27.75% (Figure 2).

Appendix C

Confusion Matrices and BOLD measurements for ROIs

For the following analyses we limit our selves to clusters of 100 voxels or more in order to ensure reliable MVPA. Specifically, to ascertain whether results were being driven by a single category, we created ROIs from our cross-decoding maps (**Table 2.1**), with the largest cluster (Precuneus, Angular Gyrus and RSC) being split into 3 subclusters (see below), bringing us to a total of 7 clusters. We computed confusion matrices for the each of these clusters for each of the two cross-decoding conditions (train on pictures, test on words; train on words, test on pictures). This was done for each subject and the mean of these confusion matrices, averaged across the two decoding conditions, are displayed in Appendix C. The column values are ground truth labels and the row values are predictions of the classifier. We used the Kruskal-Wallis test to look for accuracy differences in cross-decoding across our four categories (diagonal elements of the matrix).

We also calculated the mean and standard deviation of the normalized signal (i.e. percent signal change) for each scene category and each of the ROIs that we created to compute the confusion matrix (Appendix C). We note that these values are smaller than the beta weights used in univariate analyses because the normalized data are centered around zero. We include them only as an indication that there is no consistent mean differences among categories, underscoring the need for MVPA.

As noted above, our largest cluster encompasses a broad set of regions covering the Precuneus, the Angular gyrus, and the RSC. Because these presumably reflect different functional areas. in computing the confusion matrices we split this ROI into a smaller subset of regions. We

accomplished this by raising the statistical threshold (28.98%) on the cross-decoding accuracy map until three separate clusters emerged for the Precuneus, the Angular Gyrus and the RSC. We report the confusion matrices and the mean of the BOLD signal values for the subclusters below.

Confusion Matrix: Subcluster l-Precuneus (526 voxels; MNI :-28, -70, 46)

	Beaches	Cities	Highways	Mountains
Beaches	27.43	23.87	22.05	26.65
Cities	25.18	30.04	17.97	26.82
Highways	27.44	19.80	28.56	24.22
Mountains	24.22	21.27	19.44	35.07

Kruskal-Wallis chi-squared = 3.15, df = 3, p = 0.3686.

Confusion Matrix: Subcluster r-Angular Gyrus (239 voxels; MNI: 38, -72, 40)

	Beaches	Cities	Highways	Mountains
Beaches	29.26	20.31	22.92	27.52
Cities	25.26	27.00	23.35	24.40
Highways	24.83	17.97	33.08	24.13
Mountains	25.61	17.71	20.49	36.20

Kruskal-Wallis chi-squared = 3.11, df = 3, p = 0.3746.

Confusion Matrix: Subcluster r-RSC (102 voxels; MNI: 14, -56, 18)

	Beaches	Cities	Highways	Mountains
Beaches	29.34	19.88	21.97	28.82
Cities	25.43	26.22	21.44	26.91
Highways	24.57	24.14	24.48	26.83
Mountains	24.57	18.40	20.31	36.72

Kruskal-Wallis chi-squared = 6, df = 3, p = 0.1116.

Confusion Matrix: r-Middle Frontal Gyrus Cluster 1 (458 voxels)

	Beaches	Cities	Highways	Mountains
Beaches	28.30	24.91	21.18	25.61
Cities	24.14	28.21	22.22	25.44
Highways	25.09	19.71	30.73	24.48
Mountains	24.22	21.10	18.84	35.85

Kruskal-Wallis chi-squared = 4.61, df = 3, p = 0.2029.

Confusion Matrix: r-Middle Frontal Gyrus Cluster 2 (172 voxels).

	Beaches	Cities	Highways	Mountains
Beaches	28.74	21.62	25.26	24.39
Cities	24.40	26.56	25.96	23.09
Highways	21.35	22.83	30.64	25.17
Mountains	25.26	22.40	22.57	29.78

Kruskal-Wallis chi-squared = 1.34, df = 3, p = 0.7186.

Confusion Matrix: l-Middle Frontal Gyrus Cluster 1 (124 voxels)

	Beaches	Cities	Highways	Mountains
Beaches	24.83	25.87	28.13	21.18
Cities	22.66	25.79	29.34	22.23
Highways	23.44	24.83	32.81	18.93
Mountains	24.13	23.79	25.61	26.48

Kruskal-Wallis chi-squared = 6.16, df = 3, p = 0.1043.

Confusion Matrix**.** l-Inferior Frontal Gyrus Cluster 1

	Beaches	Cities	Highways	Mountains
Beaches	25.35	26.31	27.61	20.75
Cities	23.18	27.87	30.38	18.58
Highways	22.49	26.83	31.16	19.53
Mountains	22.23	23.70	27.87	26.22

Kruskal-Wallis chi-squared = 2.95, df = 3, p = 0.3995.

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) for the Subcluster **l-Precuneus**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	1.02E-01	7.90E-02	6.10E-02	7.94E-02	1.09E-01	1.09E-01	1.52E-01	1.04E-01
stdev	1.19E-01	8.49E-02	1.38E-01	1.12E-01	1.39E-01	9.61E-02	1.63E-01	1.12E-01

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) for the Subcluster **r-Angular Gyrus**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	1.52E-02	3.70E-05	-3.01E-02	-5.39E-02	2.68E-02	5.51E-02	7.24E-02	-6.09E-02
stdev	7.53E-02	1.17E-01	8.31E-02	1.20E-01	1.02E-01	1.22E-01	2.01E-01	1.04E-01

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) for the Subcluster **r-RSC**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	1.52E-02	3.70E-05	-3.01E-02	-5.39E-02	2.68E-02	5.51E-02	7.24E-02	-6.09E-02
stdev	7.53E-02	1.17E-01	8.31E-02	1.20E-01	1.02E-01	1.22E-01	2.01E-01	1.04E-01

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) for **r-Middle Frontal Gyrus Cluster 1**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	1.39E-02	-2.08E-02	-1.59E-02	-1.07E-02	3.04E-02	4.18E-02	6.12E-02	-4.05E-02
stdev	8.97E-02	7.57E-02	1.02E-01	7.66E-02	8.12E-02	1.10E-01	1.83E-01	7.31E-02

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) **r-Middle Frontal Gyrus Cluster 2**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	-6.23E-02	-3.88E-02	-9.77E-02	-8.54E-02	-2.92E-02	-3.23E-02	-1.06E-01	-7.01E-02
stdev	6.91E-02	7.34E-02	1.08E-01	1.03E-01	7.91E-02	1.21E-01	9.46E-02	7.64E-02

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) **for l-Middle Frontal Gyrus Cluster 1**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	1.34E-02	4.04E-02	2.92E-02	-3.38E-02	-3.52E-02	3.42E-02	-1.14E-01	-5.66E-02
stdev	1.41E-01	9.01E-02	1.39E-01	6.99E-02	1.05E-01	1.35E-01	1.74E-01	1.56E-01

The mean values of the BOLD signal for words and pictures, along with the standard deviation (stdev) for **l-Middle Frontal Gyrus Cluster 2**.

	Words				Pictures			
	Beaches	Cities	Highways	Mountains	Beaches	Cities	Highways	Mountains
Mean	1.13E-01	7.80E-02	1.32E-01	1.39E-01	3.46E-02	1.13E-01	1.83E-02	7.73E-02
stdev	1.73E-01	1.81E-01	2.35E-01	1.60E-01	7.90E-02	1.20E-01	1.94E-01	1.37E-01

Appendix D

Cross-Decoding Maps In Each Direction

The cross-decoding maps showing both directions of cross-decoding (train on pictures, test on words and train on words, test on pictures) are presented below.

Figure D.1 **A**: Cross-decoding accuracy from pictures to words (train on picture runs and test on a word run). **B**: Cross-decoding accuracy from words to pictures (train on word runs and test on a picture run). **C**: A map of the intersection of the two cross-decoding results: only pictures to words (blue), only words to pictures (orange) and regions common to both sets (red).

Figure D.1. (contd.)

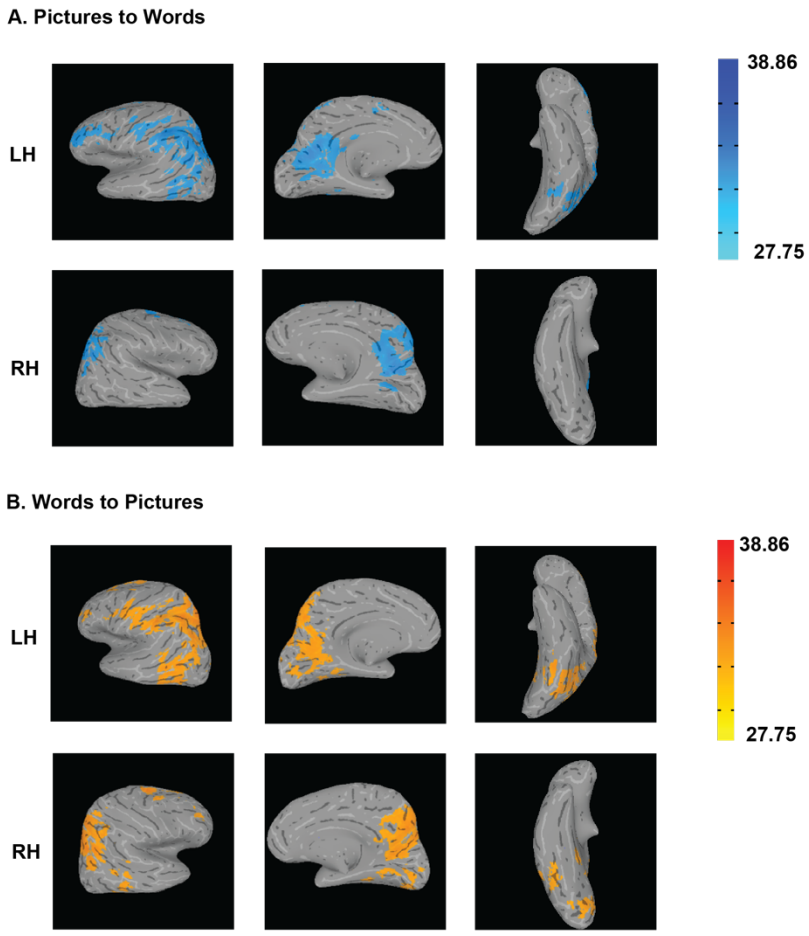


Figure D.1 (contd.)