Chapter 47

NEUROSEMANTICS AND CATEGORIES

CHRIS ELIASMITH

Contents

Abstract 1036
1. Introduction 1037
 1.1. Why “neuro”? 1037
 1.2. The explanandum 1039
2. Mental representations as neural codes 1040
 2.1. Representations 1040
 2.2. Transformation 1041
 2.3. A representational hierarchy 1042
3. The meaning of neural representations: Neurosemantics 1043
 3.1. The representation relation 1043
 3.2. A neurosemantic theory 1044
  3.2.1. Systems 1044
  3.2.2. Vehicles 1045
  3.2.3. Referents 1046
  3.2.4. Content 1047
 3.3. Discussion 1049
4. Misrepresentation 1050
5. Conclusion 1052
References 1052

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
2005 Elsevier Science B.V. All rights reserved

Abstract

A theory of category meaning that starts with the shared resources of all animals (i.e., neurons) can, if suitably constructed, provide solutions to traditional problems in semantics. I argue that traditional approaches that rely too heavily on linguistics or folk psychological categories are likely to lead us astray. In light of these methodological considerations, I turn to the more theoretical question of how to construct a semantic theory for categories informed by neuroscience. The second part of the chapter is concerned with describing such a theory and discussing some of its consequences. I present a theory of neural representations that describes them as a kind of code, and show that such an understanding scales naturally to include complex representations such as concepts. I use this understanding of representational states to underwrite a theory of semantics. However, the theory must be supplemented by what I call the statistical dependence hypothesis. Content is then determined by a combination of the states picked out by this hypothesis and the neural decoders that define subsequent transformations of the neural representations. I briefly describe a solution to the traditional problem of misrepresentation that is consistent with this theory.

1. Introduction

As described in many contributions to this volume, categories are concepts, and concepts are representations. Some claim that these representations are learned by individuals, others that they are socially constructed, and still others that they are innate. There are theories of perceptual categorization, categorization in data mining, and the metaphysics of categorization. There are contributions relating to categorization in computer science, philosophy, psychology, and neuroscience. In most such discussions, there are fundamental assumptions that usually remain implicit regarding categories. These have to do with the *meaning*, or *semantics*, of categories.

Of course, the semantics of categories has been discussed at length here and elsewhere [Loar (1981), Lycan (1984), Millikan (1984), Dretske (1988), Fodor (1998)]. However, what is unique about the discussion in this chapter is the perspective that I adopt: I address semantics from the perspective of computational neuroscience. Most attempts at addressing semantics have been linguistic or psychological and rely on more traditional accounts of computation. I argue that adopting a neuroscientific view has a number of advantages. And, to demonstrate this concretely, I outline a theory of category meaning based on recent work in computational neuroscience. To show that this theory has some advantages over past theories, I apply it to the traditional problem of misrepresentation. That is, the problem of determining how we can miscategorize entities in the world. Somewhat surprisingly, the problem of misrepresentation is simply ignored by many empirical approaches to categorization. But it is, in fact, the very central problem of determining what the appropriate labels for categories actually are.

1.1. Why "neuro"?

Many philosophers who have proposed semantic theories have focused on the propositional content of beliefs and language [see, e.g., Loar (1981), Evans (1982), Harman (1982), Lycan (1984), Block (1986), Fodor (1998)]. This project has been less than obviously successful. As Lycan (1984), a proponent of the approach, has put the point:

> Linguistics is so hard. Even after thirty years of exhausting work by scores of brilliant theorists, virtually no actual syntactic or semantic result has been established by the professional community as known. (p. 259)

But Lycan, like most researchers in the field, is determined to continue with the project using the same methods, and shunning others: "And there must be some description of this processing that yields the right predictions without descending all the way to the neuron-by-neuron level" [Lycan (1984) p. 259]. After significantly more than 30 years of difficulty, it seems rather likely that those neuron-by-neuron details actually *do* matter to a good characterization of the syntax and semantics of mental representations (and, through them, language).

There are reasons other than a simple lack of success for language-based approaches to think that neuroscience may be a better starting point for such theories. For one thing, participants in the debate generally agree that mental content should be naturalized. That is, meaning deserves a scientific explanation that refers to objects found in nature. Linguistic objects, like words, are presumably one kind of object found in nature. But, it is a mistake to give an explanation of meaning *in terms of* words and complex natural language, since this is to explain one poorly understood natural concept (mental meaning) in terms of another (linguistic meaning). In fact, such an explanation would be perfectly circular if we were giving an explanation of meaning that relied on the content-carrying capacity of words. For this reason, most language-based explanations try to introduce other notions as primitive, notions like "information," "cause," or "function." But even with these less circular primitives, assumptions regarding the structure of mental meanings (e.g., that they are language-like in various ways) are often deeply influenced by the example of natural language.

This is a serious problem because language is only a tiny domain of the application of the notion of mental meaning. Language, as most linguists understand it, is a human specialization. Thus it is unique to one species among billions. This is a good reason to think that starting with language, or even focusing on language, when constructing a theory of content is a dubious tactic. This is true unless we have prima facie evidence that most nonhuman animals do not have internal representations. We have no such evidence, and, in fact, there is significant evidence to the contrary [Redish (1999), Eliasmith (2003)]. At a minimum, the use of symbols for communication is rampant in the animal kingdom. Bee dances, monkey calls, whale songs, bird songs, etc. are all instances of animals communicating properties of the environment via symbols that refer to the things having those properties. So it is much *more* common for there to be meaning without language than meaning with language. Meaning, it seems, is prior to language.

Further support for this contention comes from the clear evidence that language is not necessary for mental meaning. People who have had the misfortune of growing up without natural language, but later learn language, are able to recall events that precede their linguistic competence [Garmon and Apsell (1997)]. So we need a theory that can account for meaning regardless of the presence of natural language.

In addition, if we take linguistic capacities to be the result of a somewhat continuous evolutionary process from less complex to more complex organisms, then the fact that language is a human specialization suggests that it is a far *more* complex phenomenon than "merely" having neural states with meaning. Even those, like Chomsky (1986), who think that language is a specifically human ability that does not have evolutionary precursors, argue that language is particularly complex. Being able to deal with the complexities of language suggests a uniquely powerful set of computational abilities in humans. To begin our explorations of meaning by examining a phenomenon found *solely* in the most complex exemplar systems with meaning is, practically speaking, a bad tactic [Bechtel and Richardson (1993)]. This might serve to explain the lack of progress noted by Lycan.

Given these considerations, it should be clear just how bad an idea it is to hang our theories of mental meaning on what we know about the structure of natural language. To claim that animals have a "language of thought" simply because our theory of meaning depends on our understanding of human language is quite clearly a claim of the tail-wagging-the-dog variety [see, e.g., Fodor (1975)]. In contrast, a theory of meaning that starts with the shared resources of all animals (i.e., neurons) can, if suitably constructed, draw the anticipated boundaries in virtue of principled differences in the use or organization of these resources. So starting at the "neuron-by-neuron" level may not be such a bad idea after all.

1.2. The explanandum

In the last 20 years or so, there has been a plethora of philosophical theories trying to provide naturalistic explanations of mental meaning. They have included covariance theories [Dretske (1981, 1988), Fodor (1981, 1998)], functional role theories [Harman (1982, 1987), Block (1986)], adaptational role theories [Millikan (1984), Dretske (1988, 1995)], and isomorphism theories [Cummins (1996), O'Brien and Opie (2004)]. AQ1 It is not always clear that these theorists have the same target in mind when providing their explanations. So, to begin, I outline what I take to be expected of a good, naturalistic theory of meaning.

For the most part, contemporary naturalistic theories of mental meaning are part of a tradition concerned primarily with linguistic meaning. In this tradition, the meaning of a sentence is the abstract proposition that the sentence expresses. For mental states, it is the thought that has this same meaning (or content). Simple sentences, like *The star is bright*, have traditionally been analyzed as an ordered pair of elements <object, predicate> whose meanings are thought to be either senses (sometimes called the "Fregean" view) or references (sometimes called the "Russellian" view).

In mental, or psycho-, semantics, however, it is not clear what the object/predicate distinction should amount to (unless one presupposes a Language of Thought, which I do not). This is because psychosemanticists have been much more concerned with the meaning of simple mental representations, i.e., categories or concepts, which, while they carry content, are often taken to be the *constituents* of sentences. Concept labels, like *dog*, do not explicitly declare objects and predicates. However, concepts are often taken to be category labels, whose members might share some common or "core" properties. And this category/property distinction is often employed like an object/predicate one (e.g., to say "My concept [DOG] includes mostly furry objects" is to say that "Dogs are furry").

For present purposes, this parallel means that doing neurosemantics is really doing semantics in the linguistic/psychological tradition [which some, like Fodor (1987), may strongly doubt]. This is because I take the meaning of a neural representation to be what that representation tells you about what it represents. An equivalent way of saying this is that the meaning of a neural representation is the set of properties ascribed to something by that representation[1]. So, I too adopt an analogous object/predicate or category/property

[1] Although I take it that, in general, the "something" can be objects, events, relations, and so on, here I will concentrate only on objects. As well, I will concentrate on predicate-object pairs only (e.g., not compound sentences).

distinction. However, this distinction can be "subpsychological" (e.g., the firing of a retinal neuron ascribes the property of a certain luminance at a retinal location). As a result, I take my explanandum to be shared with these past theories. And I take that explanandum to be: How is the set of properties ascribed to an object by its representation determined?

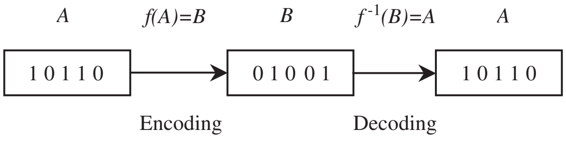
Before addressing this question directly, in Section 3, the next section takes a brief detour into recent work in computational neuroscience that sheds light on the kinds of neural states that are relevant for mental representation [Eliasmith (2003)].

2. Mental representations as neural codes

2.1. Representations

As discussed in detail in Eliasmith and Anderson (2003), it is possible to adapt a standard information theoretical account of codes to understanding representations in neural systems. Codes, in information theory, are defined in terms of a complementary encoding and decoding procedure between two alphabets. Figure 1 depicts a toy case of a code, where *B* would be said to be a code for, or representation of, *A*. In this case, both alphabets are binary digits.

In order to characterize representation in a neural system in a similar way, we need to identify each of these procedures and their relevant alphabets. The encoding procedure is straightforward: it is the transduction of stimuli by the system resulting in a series of neural action potentials. Encoding is what neuroscientists typically talk about. For example, when we present an image of a dog to a subject, some neurons or other "fire"[2]. Unfortunately, neuroscientists often stop here in their characterization of representation (i.e., they want to claim that those neurons represent a dog), but this is insufficient (stopping here is equivalent to adopting a naive causal theory). We also need to



$$f(x \in A) = \begin{cases} 0 \text{ for } x = 1 \\ 1 \text{ for } x = 0 \end{cases} \qquad f^{-1}(x \in B) = \begin{cases} 0 \text{ for } x = 1 \\ 1 \text{ for } x = 0 \end{cases}$$

Fig. 1. A simple example of an encoding/decoding relation.

[2] The precise nature of this encoding has been well quantified [Bower and Beeman (1998)].

identify a decoding procedure; otherwise, there is no way to determine the relevance of the encoding for the subject. If no information about the stimulus can be extracted from the spiking neurons, then it makes no sense to say that it represents the stimulus. After all, representations should be able to "stand in" for that which they represent. As discussed in Eliasmith and Anderson (2003), there are methods for determining good, linear decoders that decode time-varying, distributed population representations.

Having specified the encoding and decoding procedures, we still need to specify the relevant alphabets. Neuroscientists generally agree that the basic element of the encoding alphabet is the neural spike. However, there are many possibilities for how such spikes carry information. Of these possibilities, arguably the best evidence exists for a combination of timing codes and population codes [see Abbott (1994) for an overview; Salinas and Abbott (1994), Rieke et al. (1997)]. Thus, I take the encoded alphabet to be the set of temporally patterned neural spikes over populations of neurons.

Unfortunately, it is more difficult to be specific about the nature of the decoded alphabet. Nevertheless, a justifiable assumption is that what is encoded (at least initially) are physical properties in some specifiable units (e.g., m/s, kg, etc.). More complex properties (e.g., red, hot, conspecific, etc.) are inferred on the basis of representations of properties with identifiable units. As a result, the units of such properties will also be complex, and are currently undetermined (though empirically determinable).

2.2. Transformation

A representational characterization is of little use if it does not play a role in understanding the function of a neural system. Conveniently, the preceding characterization of neural representation can be easily extended to account for neural transformations. This is because, like representations, transformations can be characterized using the notion of decoding. Specifically, related methods exist to find a "transformational decoder" which extracts information *other than* what the population is taken to represent [Eliasmith and Anderson (2003)].

Given this understanding of neural computation (i.e., as encoding and (transformational) decoding), a problem arises because the information encoded in a population may now be decoded in a variety of ways. Since representation is also defined in terms of encoding and decoding, it seems that we need a way to determine which of the set of possible decodings is the relevant one for defining the "true" representation of the original population.

To resolve this problem, I have elsewhere specified that what a population represents is determined by the decoding that results in the variable that all other decodings are functions of [Eliasmith (2003)]. This does not completely resolve the difficulty since any given variable can be written as a function of another variable that is a function of it. This remaining difficulty can be resolved by noticing that the relevant variable is that variable whose units are part of a coherent and useful theory [Eliasmith (2003)]. This merely amounts to the observation that our neuroscientific theories ought to be consistent and continuous with our other physical theories. For instance, supposing that x is an object's position in the world, and that x and x^2 are decoded from some encoding in a neural population, we would claim that the population represents position because all uses of the representation can be related to position and "position" is a standard category in the sciences (whereas "position squared" is not). In other words, we should (at least initially) take cognitive systems to represent the properties that we describe the world in terms of since cognitive systems represent the world. This is not a demand for realism regarding all properties in our current descriptions of the world, but a demand for consistency between those descriptions and descriptions of cognitive systems.

2.3. A representational hierarchy

For illustrative purposes, let us consider a neural population's representation of the horizontal position of an object; such a representation is found in the lateral intraparietal cortex [Andersen, Essick and Siegel (1985)]. This representation can be understood as the encoding of a scalar variable into a series of neural spikes[3]. The units of that variable are "degrees from midline." Using the quantitative tools mentioned earlier, it is possible to determine the representational decoder. Once we have the decoder, we can then estimate what the actual position of the object is, given the neural spiking in this population. Comparing this result to an independent characterization of the object's position, we can then determine precisely how well the original property (or specific aspects of it) is represented by the neurons in the population.

This simple example not only describes how to characterize representation, it also shows how we can move from talking about single neuron responses to talking about "higher-level" variables, such as "horizontal object position." That is, we can move from discussing the "basic" representations (i.e., neural spikes) to "higher-level" representations (i.e., mathematical objects with units). As I have discussed elsewhere, even such a simple example as this demonstrates how to build up a "representational hierarchy" that permits us to move away from a "neuron-by-neuron" description, while remaining responsible to it [Eliasmith (2003)]. To continue up the hierarchy, we could talk about a larger population of neurons that encodes object position in three-dimensional space. We could then spell out the details of lower levels of the hierarchy by dissecting this higher-level description into horizontal, vertical, and depth positions, and then dissect these descriptions further into neural responses. Which description we employ will depend on the kind of explanation we need and the specific properties of the neural system we are describing (e.g., the nature of the neurons' tuning curves).

Notably, this hierarchy can be rigorously defined to include scalars, vectors, functions, and any combinations of these [Eliasmith and Anderson (2003), p. 63]. Because all of the levels of this hierarchy can be written in a standard form, it is natural to suggest that it provides a unified way of understanding representation in neurobiological systems. The wide range of mathematical objects that can be related to neural processing in this way supports the stronger claim that this hierarchy is general enough to

[3] Although this understanding is oversimplified and unrealistic, it is a useful approximation for this purpose.

capture all neural representation. That is, since all mental representations can be described as some combination of scalars, vectors, and functions, and those mathematical objects can be neurally represented using these methods, these methods can be used to describe all mental representations[4].

That being said, it is important to keep in mind the fact that there can be more than one possible implementation at each level of the hierarchy (e.g., more than one way a neural system can represent a three-dimensional vector). So this hierarchy in no way relieves us from being responsible to the empirical facts regarding a particular neurobiological system. Nevertheless, this representational hierarchy can unify our understanding of representation in neurobiological systems from neural firings to psychological-level representations.

Thus, *regardless* of what kind of mathematical objects with units higher-level representations turn out to be, the previous method for understanding neural representation applies. While adopting the representational hierarchy has definite consequences for the form of a preferred representational account, it is silent as to which particular one is correct (i.e., it does not determine which mathematical objects are appropriate for which aspects of neural function). This is a desirable result, since we simply do not know what the right higher-level representations are at this point in the history of neuroscience. While there is general agreement on the mechanism underlying individual neuron activity, there is little agreement on how populations of neurons manage to represent the world. Presumably, the right account will be the most coherent and predictively successful one – an adjudication that it is simply too early to make.

3. The meaning of neural representations: Neurosemantics

3.1. The representation relation

To this point, I have only addressed how to characterize the kinds of states that can carry content. In particular, I have done so using the tools of computational neuroscience. I have not, however, addressed how to determine the content carried by those states.

Standard accounts of the representation relation, including both causal and conceptual role theories, identify a three-place relation in describing representation: for causal theories, there is the representation, the thing it represents, and the context under which it is a representation (and not just an effect)[5]. For conceptual role theories, there is the representation, the thing it represents, and the role it plays (i.e., its context as defined by the system of concepts).

[4] This may seem implausible if we are concerned that canonical cognitive representations such as images and sentences are not mathematical objects of these kinds. However, note that any digitizable representation can be considered as a vector, and functions naturally account for analog representations.

[5] For Fodor, such a context is defined by the nomic relations that obtain, in particular, those that have asymmetric dependencies and those that do not. For Dretske, the context is determined through an evolutionary and learning history.

One of the reasons that this three-place relation is so ubiquitous stems, I suspect, from the nearly universal commitment among philosophers and neuroscientists to understanding neurobiological systems as information processing systems [see, e.g., Dretske (1981), Bialek and Rieke (1992), Van Essen and Anderson (1995), Rieke et al. (1997), Fodor (1998), Koch (1998), Eliasmith and Anderson (2003)]. Formally, the information relation is a three-place relation: A {channel} carries {information} with respect to a {receiver} [Reza (1994) p. 2]. These three places are necessary and sufficient for an adequate and general definition of Shannon and Weaver-style information [Shannon (1948/1949)]. Also notice that they loosely align with the three places of the representation relation as described above (i.e., channels or vehicles carry information or content with respect to receivers or systems). It is not surprising that both the representation relation and the information relation are three-place relations since the nature of the latter often informs intuitions about the nature of the former, given such a commitment.

However, the metaphor relating mental representation and information does not hold up under scrutiny. In particular, as Dretske (1983) has been at pains to point out, there is no such thing as "misinformation" in the same sense as there is "misrepresentation." That is, information (in the technical sense) is never wrong about anything. Representations, by contrast, can be. This disanalogy is important because it highlights the need to identify a fourth element in the representation relation: the *referent*, that is, the object or event that the content of the representation is supposed to be about. The importance of this element, of course, is something like what Frege wanted to highlight with his distinction between reference and sense (although it is not the same).

I take it, in fact, that failure to distinguish contents and referents is what leads to many of the difficulties for both causal and conceptual role theories. Both equate contents and referents, though by slightly different routes. Specifically, causal theories take referents to be contents, with the result that surreptitiously changing referents leads to sometimes counterintuitive shifts in meaning and difficulties explaining misrepresentation. Conceptual role theories, in contrast, are in a position where it seems they must take contents to be referents, not allowing them to explain the relevance of truth conditions to meaning.

Consequently, I adopt the following *four*-place schema for the representation relation (the "representation schema"):

A {vehicle} represents a {content} regarding a {referent} with respect to a {system}.

3.2. A neurosemantic theory

Given this representation schema, what a semantic theory must do is explain what the four constituting theoretical objects are. The following description is based on work presented in Eliasmith (2000).

3.2.1. Systems

Although I call the fourth relatum the "system", as I take this formulation of the representation relation to be general, philosophers of mind tend to narrow the scope of this schema to include only natural biological systems (often just human beings). In other words, they are not interested in representation writ large, but only in neurobiological representation. I adopt this narrower perspective as well. So, the fourth element of the schema will be, for my purposes, "the (complete) nervous system." This is in contrast to understanding "system" as either a more general physical system or an abstract representational system [see, e.g., Goodman (1968)].

3.2.2. Vehicles

Vehicles are the internal physical objects, or "representations," that carry representational contents. As evident from Section 2.3, I take there to be two kinds of vehicles: basic and higher level. Basic vehicles are neurons, understood as functional units. Higher-level vehicles are simply sets of neurons.

To get a better understanding of what this distinction entails, consider a debate in neuroscience regarding whether neurons in visual area V1 are "edge detectors" or "orientation filters" [Van Essen and Gallant (1994)]. This example raises a concern about the distinction I have drawn between basic and higher-level vehicles. Notice that the terms "edge detectors" and "orientation filters": (1) are descriptions of single neurons; and (2) pick out vehicles based on their contents.

This first point raises the concern that taking basic vehicles as uncontentiously identifiable with single neurons is simply mistaken. Since there are debates over such terminology, it seems clear that the identities of basic vehicles really are not uncontentious after all. However, I take it that neurons are basic vehicles only as *functional units*, not as carriers of content. That is, it has been well established that single neurons have certain physical properties that result in predictable system-independent functioning. The plethora of highly detailed single-cell models attests to this fact [e.g., Bower and Beeman (1998)]. However, it is not well established what the behavior of such single neurons means. This leads us to the second point.

The reason that the identity of vehicles is not easy to settle is that vehicles are often named after the contents they supposedly carry. "Edge detectors" are so called because they are thought to be activated by edges and used to detect edges. "Orientation filters" are so called because they are thought to be activated by spatial orientations and used to analyze orientation gradients in visual images. In some ways, this seems unnecessarily confusing, but it demonstrates how important content determination is to theories in neuroscience as well as philosophy. In any case, once we keep in mind that the naming practices of vehicles have a tendency to rely on content claims, claims about the nature of vehicles (*simpliciter*) become less contentious. So, saying that basic vehicles are neurons as functional units and higher-level vehicles are sets of neurons should not be taken as saying too much.

However, it also means that claims about specific higher-level vehicles (i.e., the precise set of neurons that answers to a specific content claim) rely largely on our content claims. Saying, of a neuron, that it is an "edge detector" is saying that it instantiates a higher-level vehicle (i.e., it is a set of one neuron) that can carry content about edges.

To adjudicate this claim, we need to know what it is for a vehicle to carry content (as well as how well this claim fits into an overall theory of neural processing). Before discussing content, however, let me address referents.

3.2.3. Referents

Referents are the external objects that representations assign properties to. One traditional way of determining the "reference" (a closely related, but distinct notion) of a representation is to rely on causes. However, as Dretske (1981, pp. 26–33) rightly notes, a theory of cause does not tell you which causes are important for representational content – which, in this case, means which causes are referents.

So the relevant causes need to be somehow "specially" related to the vehicles that are carrying content about them. That is, vehicles carry contents about things that they are, in some sense, good at carrying information about. So I need to say both what this special relation is, and what causes are, in order to determine what will count as the referent of a vehicle.

Of course, discussions of the nature of cause can be lengthy enough in themselves. Here I simply adopt David Fair's (1979) approach, which identifies token causes with the transfer of energy momentum ("energy" from now on)[6]. For example, gas tanks cause gas gauges to register gas levels because there is a transfer of energy between gas tank levels and gas gauges.

AQ2

More specific to semantic theories is the question of the "specialness" of the relation between the referents and the vehicles. Notably, both the referents, *r,* and the vehicles, *v,* can be measured in their own relevant units (i.e., the decoded alphabet and encoded alphabet, respectively). Many aspects of the relation between two such measurable quantities is determined by their joint probability distribution, $p(r,v)$. This distribution can be used to determine to what degree the quantities are statistically dependent. If two quantities are dependent, then there is some relation or other between changes in one and changes in the other. A good way to quantify the degree of statistical dependence is to rely on a measure called "mutual information." As a result, I use the two terms interchangeably[7].

In fact, I think that mutual information is precisely the quantity we need to underwrite referent (and content) claims. Specifically, let me propose a preliminary version of the "statistical dependence hypothesis:"

> The referent of a vehicle is the set of causes that has the highest statistical dependence with the vehicle under all stimulus conditions.

This hypothesis makes two important assertions: (1) the highest statistical dependency picks out referents; and (2) referents are *causes*, i.e., they have an energy transfer

[6] Similar theories are suggested by Aronson (1971), Castaneda (1984), and Strawson (1987).

[7] A different, but related, semantic theory that employs mutual information measures can be found in Usher (2001).

with the relevant vehicles. Cause plays a central role here as it rules out "accidental" statistical dependencies (e.g., the clock on the Peace Tower striking 12 and my going to lunch in Montreal). In general, statistical dependencies are too weak to properly underwrite a theory of content on their own.

Despite incorporating causes, this preliminary version of the hypothesis is still inadequate. In particular, it results in a kind of solipsism. This is because the highest dependency of any given vehicle is probably with another vehicle that transfers energy to it, not with something in the external world.

Fortunately, this is a rather artificial problem. Theories of representation are precisely theories that describe the relation between a representational system and what it represents. As a result, my previous identification of the system can be invoked here to limit the application of the statistical dependence hypothesis. That is, we only apply such a referent-determining rule external to the system. This results in a modified, and final, version of the statistical dependence hypothesis:

> The referent of a vehicle is the set of causes that has the highest statistical dependence with the neural responses under all stimulus conditions and does not fall under the computational description.

The computational description here refers to the account of neural functioning provided by the theory of neural representation and computation discussed earlier. This description is adequate for only the behavior of neurons; that is, the components of the system.

A remaining concern with this formulation of the statistical dependence hypothesis pertains to the concept of "stimulus conditions." There are a number of options for dealing with stimulus conditions, but let me just consider one here: appeal to other scientific theories to individuate stimulus conditions. Thus, stimulus conditions change if any physical variable relating the referent and the representation changes. Such variables include things like distance, illumination, relative velocity, etc. However, this definition makes it difficult to distinguish incremental changes in these more intuitive stimulus condition variables from incremental changes in variables that specify the physical properties of the referent itself. This is not a problem for the theory I have presented, although it may be a problem for determining what the referent is (i.e., in answering the question "is it a dog or a cat?"). Some such "stimulus conditions" would be difficult to realize because we cannot construct "dogcats." That is, we cannot "morph" dogs into cats in the real world. This may limit our ability to systematically examine "all possible" stimulus conditions. Although this may initially seem problematic, the real-world development of an animal's conceptual categories would also be so limited – so it may not be a hindrance after all.

3.2.4. Content

Recall that I take content to be the properties ascribed to a referent by a vehicle. Recall as well that the theory of neural representation outlined earlier can be expected to play a central role in content determination. However, as discussed, it does not by itself tell us how to make content ascriptions. This is because it is strictly a theoretical or computational description of the functioning of neural systems. In order to inform the theory of content, this computational description needs to be related to a semantic one.

Recall that a neural representation is defined by identifying an encoding and decoding process. Previously, I mentioned that there are methods for finding the representational decoder that allow very good reproduction of the encoded signal. In other words, the decoder tells us how to relate the neural signal to the encoded signal, which means it tells us what properties of the encoded signal are "saved" by the neural signal. This in turn tells us which properties are ascribed to a referent by neural activity.

So content is determined by the decoders. But how are the decoders determined? Simply put, they are found by minimizing the difference (an "error" or "energy") between the signals being represented and the decoded neural activities over all represented input signals. This tells us that the decoders depend on the relation between two things: (1) the signals to be represented, *s*; and (2) the neural activities, or response, *r*. The response, *r*, is the activity of a vehicle, and the signals, *s*, are the referent[8].

In effect, then, the decoders describe a rule that determines what properties the current neural activities in a population of neurons ascribe to the current referent of the population. That rule is determined by examining the statistical dependence the vehicle has with a referent over all stimulus conditions.

This description of content highlights an important distinction between two kinds of content, what I will call "occurrent" content and "conceptual" content. The former applies to current neural activity and the current referent, under the current stimulus conditions. The latter applies to the determination of the decoders, over all stimulus conditions. Noticing this difference makes it clear that the statistical dependence hypothesis is applicable for the determination of conceptual content. To extend this application to occurrent content, let me suggest the following corollary:

> The occurrent referent of a vehicle is the set of causes that has the highest statistical dependence with the neural responses under the stimulus conditions in which it occurs.

Thus, the process of content determination stays the same, but referent determination varies across these two kinds of content. I will revisit this distinction in more detail in my account of misrepresentation. For now, it merely needs to be clear that the statistical dependence that holds over all stimulus conditions can be quite different from that that holds under a particular subset of those stimulus conditions. According to this theory, such a difference in statistical dependence results in a difference in referent. The properties ascribed (i.e., occurrent content) to such a referent by neural activity are determined by the decoders, which are themselves determined by property ascriptions under all stimulus conditions (conceptual content).

[8] A technical discussion of how these are combined can be found in Eliasmith and Anderson (2003).

3.3. Discussion

Because statistical dependence plays such an important role in this theory, it is worthwhile discussing it in more detail. Adopting statistical dependence as a means of referent determination has a number of beneficial consequences. First, statistical dependence comes in degrees. The highest dependence of one vehicle with its referent may be higher than the highest dependence of another vehicle with its referent. The strength of the dependence maps nicely onto the precision of the representation in question. If, for example, the occurrent dependence of my representation of a dog's position with the dog's actual position is nearly perfect (which means any changes in my relation to the dog are reflected in changes in the contents of my vehicle), we know that my representation is precise.

A second benefit of using statistical dependence is that it can help minimize our reliance on intuitions about what is represented by cognitive systems. In other words, it provides a means of systematically examining the features of the environment on which the vehicle is statistically dependent; we will be in a position to discover, not stipulate, what the referent of the vehicle is. In other words, it helps relieve us of our linguistic bias, which has informed most past theories of categorization.

If we have a set of neurons that we believe to be a vehicle, we can explicitly construct the joint probability histogram between those neurons and features of the environment and thus discover the referent of that vehicle. In other words, we can test our hypotheses about possible vehicles. If we suppose that a certain set of neurons acts as a "dog" vehicle, we can test that hypothesis by seeing how statistically dependent the vehicle and referent are. If they have no dependencies, or have better dependencies with other vehicles or referents, our hypothesis about the nature of the vehicle will change. The converse may occur as well: what we take to be good referents will help us determine vehicles. These methodologies are complementary and serve to give us a good idea of what vehicles and referents there are. What is most important is that this theory allows us to make *discoveries*; no doubt we will be surprised by some of the things biological systems depend on for representing the environment.

The last consequence of adopting statistical dependence I will discuss is a subtler one. Because referents are related to higher-level vehicles by statistical dependencies, our identification of referents and of higher-level vehicles is mutually dependent. Furthermore, referents help determine contents, so contents depend on referents as well. Therefore, content depends on properties of vehicles. This should not be too surprising. If you know all there is to know about a vehicle, you will know its possible content. This is not to say that the vehicles we discover will *determine* content; rather, they will *constrain* possible content. Psychophysics, for example, tells us that there are certain dynamic ranges over which our retinal cells can encode brightness. Outside of those ranges, differences in intensity are indistinguishable; that is a fact about physiology. If the vehicles cannot carry such differences in intensity, then those differences cannot be used by the system to react to the environment; consequently, those vehicles cannot carry content about those differences.

What this means, then, is that vehicles and contents are not independent. In a more traditional turn of phrase: syntax and semantics are not independent[9]. The stuff that carries meaning (syntax/vehicles) also helps to determine meaning (semantics/content). As powerful as natural languages are, they are stuck with certain vehicles. There are contents that these vehicles carry well and there are contents they do not carry well. A fixation on language for understanding categorization is likely to bias our consideration of possible categories, potentially blocking the route to a successful theory.

4. Misrepresentation

Dretske (1995) notes that the problem of misrepresentation is above all the problem of explaining how we assign the wrong properties to an object [Dretske (1994) p. 472, (1995)]. This is the same as what Fodor (1987) calls the "disjunction problem." It is the problem of explaining how a representation can assign a specific set of properties to a referent even though the representation can be caused by a disjunction of referents that do not all have those properties. For example, according to a naive causal theory if my "dog" vehicle is caused by a cat under some circumstances, that should mean that the vehicle is actually a "dog or cat" vehicle (since it can be caused by either). So the problem is how do I explain ascribing the property "dog" to a cat in certain cases (rather than just tokening a disjunctive vehicle)? How, in other words, can I explain when my categorizing a cat as a dog is a *mistake*. In this section, I address this problem and show how the theory I have presented can avoid it.

Philosophers tend to presume that cases of representing fall into one of two categories: right or wrong. This, it seems to me, is a mistake. Dretske (1994 p. 472), for example, defines misrepresentation as saying of something that does not have a given property, that it has that property; for example, saying that a black dog is brown. In other words, misrepresentation is representing one thing (a black dog) as another (a brown dog). Unfortunately, under a strict application of this definition, it is not clear that we ever get anything right – that we *ever* represent.

Consider my representing a black dog standing in front of me. Suppose that after 3 min the dog is removed from my sight. I would then be able to answer all sorts of questions about its shape, size, length, etc., based solely on my representation of the dog. However, each of my answers would be inaccurate in some way. Consider the line of questioning: What color was the dog? Answer: Black. Dark or light black? Answer: Dark black. This color (showing a color swatch)? Or this color (another swatch)? etc. There is little doubt that I would eventually answer incorrectly. Does that mean that I am attributing to this dog a property it does not have (i.e., a certain shade of black)? Yes, it does. Does it mean that we should say I am misrepresenting the dog? According to

[9] For arguments concerning the benefit of blurring the distinctions between vehicles and contents, see, e.g., Langacker (1987) and Mohana and Wee (1999). See Eliasmith and Thagard (2001) for an example of how blurring the distinction may help explain the nature of high-level cognitive processes

Dretske's definition it does, but I think that such an answer is overly hasty. In order to say why, consider some elementary distinctions in measurement theory.

Measurements are said to be *accurate* if they are near the right value. If I measure darkness and get the right answer, I have made an accurate measurement. Measurements are said to be *precise* if they are reproducible. If I measure the darkness of a color swatch over and over again, and get the same answer every time, I am making precise measurements of darkness. If my measurements are precise and accurate, they are said to be *exact*. Notably, precision is a property of a set of measurements, while accuracy is a property of a single measurement. But we can define the accuracy of a *set* of measurements as the *average* nearness to the right value. In statistical terms, precision is measured by the variance of a set of measurements, while accuracy is the difference between the average measurement and the correct answer.

What the line of questioning above is doing, is probing the representer with increasing degrees of precision. Although the representer may be perfectly accurate at one degree of precision (black versus white) the representer may be inaccurate at another (one color swatch versus another). Neural representations have a limited degree of precision; only about three bits of information are transmitted per spike [Rieke et al. (1997)]. So we should not be surprised that we will *eventually* misascribe properties; it is not possible to be consistently accurate to a degree of precision greater than our "measuring device" can provide.

What is important here is that representations are best characterized as "better" or "worse," not "right" or "wrong." "Better" means high accuracy with high degrees of precision. "Worse" means low accuracy with low degrees of precision. We may want to make "right" and "wrong" claims at a given level of precision, but in doing so we would have to make an argument as to why being below some standard of accuracy at a given precision is a good criterion for making this distinction. Dretske's definition clearly does nothing of the sort. Using the term "misrepresentation" to divide representations into two groups obscures important subtleties of representation. We will have a more general understanding of misrepresentation (i.e., one that does not depend on choosing particular standards and can account for degrees of deviation from any given standard), if we accept that representations come in degrees, i.e., that they lie on a continuum from good to bad.

The preceding discussion lays the groundwork for addressing Fodor's disjunction problem head on. The challenge of the disjunction problem, then, is to explain how my representation of something (a cat) could mean to me that it had some property (the property of being a dog) even though it was caused by something that I would say did not have that property (a cat). The solution is that we can explain this case of misrepresentation by a careful application of the statistical dependence hypothesis. More specifically, we can notice that the hypothesis and its corollary give different answers. In particular, under all stimulus conditions, this vehicle has the highest statistical dependency with dogs even though, under this condition, this vehicle has the highest dependency with a cat.

But, does this really solve the *disjunction* problem? Will it not be the case that the highest statistical dependency holds between dogs-or-this-cat under all stimulus conditions?

In fact, no. This cat under *all* stimulus conditions will not have a high statistical dependency with my "dog" vehicle; it will have the highest dependency with my "cat" vehicle. It is only under *this* stimulus condition that it has a high statistical dependency with my "dog" vehicle. In other words, because there is another vehicle (the "cat" vehicle) that has a higher dependency with this referent under all stimulus conditions, it cannot be this vehicle (the "dog" vehicle) that has this cat as its referent. Note also that this solution is possible because the notion of representation/misrepresentation is a graded one.

We can introduce a further variant in order to try to preserve the disjunction problem. That is, perhaps the problem is that my vehicle has the highest statistical dependence with dogs-or-this-cat-under-these-conditions. Luckily, this disjunction can be ruled out because it includes a specification of stimulus conditions. We cannot find a dependence between a vehicle and something-under-a-stimulus-condition under all stimulus conditions since most of the stimulus conditions are ruled out by such a characterization. It would be self-contradictory to try and determine such a dependency. Thus, I take it that the account I have given of content determination can satisfactorily account for misrepresentation.

5. Conclusion

The theory of mental meaning I have presented begins with considerations of representation in neural systems. For this reason, I have called it a theory of neurosemantics. However, because I have identified explicit relations between neural and higher-level representations, I take the theory to be applicable to the more traditional problems of psychosemantics. I have shown how the referent relation, and the statistical dependence hypothesis on which it relies, can be used to address at least one such problem, that of misrepresentation. There are, of course, many issues which I have not addressed in this paper, but I take this kind of initial success to bode well for neurosemantics in general.

Finally, the fact that I have successfully adopted a neuroscientific view of the problem of naturalizing semantics demonstrates that, indeed, the "neuron-by-neuron" account so scorned by Lycan and others may turn out to be the most fruitful one. The irony that trying to understand categorization often requires us to divest ourselves of our own cherished categories can best be embraced by the adoption of quantitative tools applied to characterizations of the system we are trying to understand. At the moment, neuroscience (combined with engineering), more than linguistics, philosophy, or psychology, provides the kinds of descriptions of cognitive systems that can inform such a characterization. Getting to the heart of categorization is a matter of properly understanding the biological implementation of cognitive systems: once again, the devil is in the details.

References

Abbott, L.F. (1994), "Decoding neuronal firing and modelling neural networks", Quarterly Review of Biophysics 27:291–331.

Andersen, R.A., G.K. Essick and R.M. Siegel (1985), "The encoding of spatial location by posterior parietal neurons", Science 230:456–458.

Aronson, J. (1971), "The legacy of Hume's analysis of causation", Studies in the History and Philosophy of Science 7:135–136.

Bechtel, W., and R.C. Richardson (1993), Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research (Princeton University Press, Princeton, NJ).

Bialek, W., and F. Rieke (1992), "Reliability and information transmission in spiking neurons", Trends in Neurosciences 15:428–434.

Block, N. (1986), "Advertisement for a semantics for psychology", Midwest Studies in Philosophy 10:615–678.

Bower, J., and D. Beeman, (1998), The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System (Springer, New York).

Castaneda, H. (1984), "Causes, causity, and energy", Midwest Studies in Philosophy 9:17–28.

Chomsky, N. (1986), Knowledge of Language (Praeger, New York).

Cummins, R. (1996), Representations, Targets, and Attitudes (MIT Press, Cambridge, MA).

Dretske, F. (1981), Knowledge and the Flow of Information (MIT Press, Cambridge, MA).

Dretske, F. (1983), "Precis of "Knowledge and the flow of information", Behavioral and Brain Sciences 6:55–63.

Dretske, F. (1988), Explaining Behavior (MIT Press, Cambridge, MA).

Dretske, F. (1994), "If you can't make one, you don't know how it works", Midwest Studies in Philosophy 19:615–678.

Dretske, F. (1995), Naturalizing the Mind (MIT Press, Cambridge, MA).

Eliasmith, C. (2000), "How neurons mean: A neurocomputational theory of representational content", Ph.D. dissertation, Washington University, St. Louis.

Eliasmith, C. (2003), "Moving beyond metaphors: Understanding the mind for what it is", Journal of Philosophy 100:493–520.

Eliasmith, C., and C.H. Anderson, (2003), Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems (MIT Press, Cambridge, MA).

Eliasmith, C., and P. Thagard, (2001), "Integrating structure and meaning: A distributed model of analogical mapping", Cognitive Science 25:245–286.

Evans, G. (1982), Varieties of Reference (Oxford University Press, New York).

Fair, D. (1979), "Causation and the flow of energy", Erkenntnis 14:219–250.

Fodor, J. (1975), The Language of Thought (Crowell, New York).

Fodor, J. (1981), Representations (MIT Press, Cambridge, MA).

Fodor, J. (1987), Psychosemantics (MIT Press, Cambridge, MA).

Fodor, J. (1998), Concepts: Where Cognitive Science went Wrong (Oxford University Press, New York).

Garmon, L. (Producer and Director), and P. Apsell (Executive Producer) (1997), "Secret of the wild child [Television series episode]", in: Nova (Public Broadcasting Service, Boston, MA).

Goodman, N. (1968), Languages of Art, (Hackett Publishing Company, Indianapolis, IN).

Harman, G. (1982), "Conceptual role semantics", Notre Dame Journal of Formal Logic 23:242–256.

Harman, G. (1987), "(Nonsolipsistic) conceptual role semantics", in: E. LePore, ed., Semantics of Natural Language (Academic Press, New York) 55–81.

Koch, C. (1998), Biophysics of Computation: Information Processing in Single Neurons (Oxford University Press, Oxford).

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

Loar, B. (1981), Mind and Meaning (Cambridge University Press, London).

Lycan, W. (1984), Logical Form in Natural Language (MIT Press, Cambridge, MA).

Millikan, R. G. (1984), Language, Thought and Other Biological Categories (MIT Press, Cambridge, MA).

Mohana, T., and L. Wee (1999), Grammatical Semantics: Evidence for Structure in Meaning (CSLI Publications, Stanford, CA).

O'Brien, G. and J. Opie (2004), "Notes towards a structuralist theory of mental representation", in: H. Clapin, P. Staines and P. Slezak eds., Representation in Mind: New Approaches to Mental Representation (Elsevier, Amsterdam).

Redish, D. (1999), Beyond the Cognitive Map (MIT Press, Cambridge, MA).

Reza, F.M. (1994), An Introduction to Information Theory (Dover, New York).

Rieke, F., D. Warland, R. de Ruyter van Steveninck and W. Bialek (1997), Spikes: Exploring the Neural Code, (MIT Press, Cambridge, MA).

Salinas, E., and L. Abbott (1994), “Vector reconstruction from firing rates”, Journal of Computational Neuroscience 1:89–107.

Shannon, C. (1948/1949), “A mathematical theory of communication”, in: C. Shannon and W. Weaver, eds., The Mathematical Theory of Communication (University of Illinois Press, Urbana, IL) 623–656.

Strawson, G. (1987), “Realism and causation”, Philosophical Quarterly 37:253–277.

Usher, M. (2001), “A statistical referential theory of content: Using information theory to account for misrepresentation”, Mind and Language 16:311–334.

Van Essen, D.C., and C. H. Anderson (1995), “Information processing strategies and pathways in the primate visual system”, in: S.F. Zornetzer, J.L. Davis and C. Lau, eds., An Introduction to Neural and Electron Networks (Academic Press, Orlando, FL) 45–76.

Van Essen, D., and J. Gallant (1994), “Neural mechanisms of form and motion processing in the primate visual system”, Neuron 13:1–10.