A neuropsychological theory of metaphor

Marc L. Schnitzer *, Mark A. Pedreira

Universidad de Puerto Rico, P.O. Box 23356, San Juan 00931-3356, Puerto Rico

Accepted 8 October 2004

Abstract

In their connectionist approach, Schnitzer and Pedreira offer an explanation for the existence of metaphor, rejecting 'process' and 'representation' accounts for a more 'vehicular' theory that grounds the study of metaphor in the way the nervous system works. From their perspective, most contemporary studies of metaphor are process oriented in that they study how metaphor works and its utility for communication rather than the neuropsychological nature of metaphor. They discuss some of the issues raised by cognitivist metaphor theorists, particularly concerning the conceptual nature of metaphor. The authors then present pedagogical illustrations concerning neural networks that demonstrate not only the metaphorical nature of thought but also the importance of such networks in learning. Their neural-network proposal can account for phenomena as simple as fruit words and concepts and as complex as metaphorical expressions about emotions. They conclude by demonstrating that a connectionist approach can resolve some important dilemmas in cognitivist theory.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Cognition; Connectionism; Metaphor; Neural network; Neuropsychology; Structural realism

1. Introduction

In this paper we address the issue of why metaphors exist—why one should wish to express something in terms of something else. What we present is not a functional
account. That is, we are not looking for the answer to this question in terms of communicative needs or desires or even in terms of the utility of metaphor in the organization of thought. In the words of Richards, who might be seen as an early pioneer in the cognitive study of metaphor, ‘Thought is metaphoric, and proceeds by comparison, and the metaphors of language derive therefrom’ (Richards, 1936, p. 94). What we hope to present herein is an explanation of why thought should be metaphoric, given the way that the human mind–brain is constructed. We intend to couch this explanation in terms of a theory of brain mechanism known as connectionism. Conversely, the ability of connectionism to explain why metaphors should exist indirectly lends support to the theory’s model of the mind–brain as it relates to thought and language.

Contemporary analyses of metaphor may generally be thought of broadly as being of either the ‘Classical’ or the ‘Cognitivist’ approach (see, for example, Saeed, 2003), based on whether one is willing to countenance the existence of literal meanings (i.e., the classical literal/figurative distinction), or whether one rejects the notion of literal meanings in favor of the view that all language is metaphoric. The classical approach regards metaphors as something of an anomaly, as usages that deviate from ‘standard usage’, or at least as relatively uncommon usages. Johnson (1981) summarizes the traditional view as follows: ‘A metaphor is an elliptical simile useful for stylistic, rhetorical, and didactic purposes, but which can be translated into a literal paraphrase without any loss of cognitive content’ (p. 4). For most of the history of Western thought on metaphor, it has been considered at best a deviant use of language to adorn speech and achieve eloquence (e.g. Cicero, Quintilian, and Longinus) or at worst a misuse of words in order to deceive (e.g. Plato, Hobbes, and Locke).

2. Modern approaches to metaphor

Gibbs (1994) provides a thorough taxonomy of contemporary approaches to metaphor, which accounts for some of the major scholarship on metaphor during the

---

1 Cicero, Quintilian, and Longinus, following Aristotle’s discussion of metaphor as ‘unusual’ or ‘exotic’ diction (‘diverge[nce] from the standard’), discuss the motivations for metaphor as copiousness, clarity, amplification, and sublimity; in classical theory, this trope achieves eloquence at its best, a ‘frigid’ style at its worst (see especially Aristotle, Poetics 1459a [Halliwell, 1995], Rhetoric 1405a8-10 [Kennedy, 1991] Cicero, De Oratore 3.39-40 [Sutton and Rackham, 1942]; Quintilian, Institutio Oratoria 8.6.4-18 [Butler, 1921]; Longinus, On the Sublime 30-32 [Fyfe and Russell, 1995]).

Plato (1968, 1994, 2002), Hobbes (1660), and Locke (1690), however, focus almost exclusively on the potential of figurative language in general, and metaphor in particular, to deceive—the most common causes of deception brought about by the ‘abuse of words’ (rhetoric’s ‘perfect cheat’, as Locke (1690) famously calls it in An Essay concerning Human Understanding, Book 3, Chapter 10) or by sophistic political motivations (recall Plato’s attack on sophists in the Phaedrus [2002 translation] and Gorgias [1994 translation] and his proposed banishment of orators and poets in The Republic [1968 translation]).
last half century. To frame our own discussion, we briefly summarize each position and refer the reader to Gibbs for detailed discussion and criticism.

1. Substitution. A metaphor is nothing more than an ornamental device and can be replaced by a literal term without loss. An example would be *Man is a wolf* substituting for *Man is fierce* (Black, 1955). Metaphors are useful for coining terms for new concepts.

2. Comparison. A metaphor is a condensed statement of a similarity or analogy. The utterance of *The lion roared* in response to a query regarding the boss's mood would involve comparing the boss's display of anger to a lion's roaring. Some theories involve the transfer or cancellation of incompatible semantic features between the tenor and the vehicle, e.g. *Their legislative program is a rocket to the moon* (Cohen, 1979), in which features of rocket incompatible with legislative program are cancelled.

3. Meaninglessness. Davidson (1979) has taken the position that a ‘metaphor means what the words mean and nothing more’ (30). What makes a metaphor different from a literal statement is that it is used to suggest something novel. When one utters *John is a pig*, one is making a literal statement about John. But the speaker may use this sentence to intimate something else about John's character or appearance.

4. Anomaly. Metaphor is identified with a violation of selection restrictions between the tenor and the vehicle. *The stone died* is semantically anomalous because the verb *die* requires an animate subject.

5. Speech acts. Searle (1979) has proposed an approach in which the notion of metaphor does not pertain to words or sentences per se, but to what a speaker means when (s)he utters words or sentences (presumed to have 'literal' meanings). This is similar to position #3, except that the latter would deny the existence of this speaker’s ‘utterance meaning.’

6. Loose talk. According to this view, based in ‘relevance theory’ (Sperber and Wilson, 1986; Wilson and Sperber, 2002) metaphors are examples of ‘loose talk,’ which may often be the most effective way of achieving ‘optimal relevance’ in an utterance. Thus, for example, when stating the price of a house, it is not usually expected that one states the final price to the penny or even the nearest dollar; more likely, the price would be rounded off to the nearest thousand dollars, a more relevant characterization than the more precisely correct one. Thus *My neighbor is a dragon* is a loose way of saying that my neighbor is ‘fierce and unfriendly’ in a more relevant way.

7. Interaction. Black (1955) introduced the idea that in the metaphor *A is B*, the ‘system of associated commonplaces’ of A interacts with that of B to ‘create’ metaphorical meaning. We use one system of commonplaces to ‘filter’ our conception of some other system. Thus, in *Man is a wolf*, we use the system of commonplaces (true or not) regarding *wolf* to filter that of *man*, thereby arriving at the metaphorical interpretation. In his discussion, Gibbs considers four psycholinguistic models that have been proposed to explain how a metaphor ‘creates’ new meanings.
7a. Salience imbalance. In similarity statements, the second term (in a metaphor, the vehicle) should show significantly more salience of the attributes associated with the terms. For example, in *Sermons are like sleeping pills* the attribute of ‘inducing drowsiness’ is much more salient for sleeping pills than for sermons.

7b. Domains interaction. Metaphorical interpretation requires assessment of similarity both between and within domains. Between-domain similarity refers to the degree in which the tenor and vehicle resemble each other. Within-domain similarity refers to the extent that the characteristics of the tenor and vehicle domains are structurally parallel, which would require the domains of both the tenor and the vehicle to be scaled the same way.

7c. Structure mapping. Mapping of knowledge between the tenor and the vehicle is based on analogies (which ignore object properties and focus on relational structure), appearances (which ignore relational structure and focus on object properties), or literal similarities (which focus on both). Examples given for each of these are, respectively, *Look, he’s winding up the watch of his wit; by and by it will strike* (analogy), *Her arms were like twin swans* (appearance), and *Plant stems are drinking straws for thirsty trees* (literal similarity).

7d. Class inclusion. Metaphors are class inclusion statements. In *My job is a jail* the tenor my job is assigned to the category of the vehicle, jail, which is a newly created ad hoc category and ‘at the same time is a prototypical example of that category.’ (Gibbs, 1994, p. 247).

8. Conceptual structure. ‘Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature’ (Lakoff and Johnson, 1980, p. 3). The world is experienced in terms of ‘experiential gestals’ such as ARGUMENT IS WAR, LIFE IS A JOURNEY, LOVE IS WAR. Lakoff and Johnson (1999, pp. 50–54)) present a list of what, following Grady (1997), they call ‘primary metaphors’ (a.k.a. ‘basic metaphors’) which, ex hypothesi, arise ‘naturally, unconsciously, through everyday experience’ (p. 46), and which form the basis for complex metaphors. Among them are IMPORTANT IS BIG, MORE IS UP, PURPOSES ARE DESTINATIONS, KNOWING IS SEEING, UNDERSTANDING IS GRASPING, CATEGORIES ARE CONTAINERS, and BAD IS STINKY.

This last approach considered by Gibbs, and the one which he defends, is also known as the cognitivist approach, and has been strongly defended by proponents of cognitive linguistics such as Lakoff, Johnson, and Turner, as well as Gibbs. Proponents of this approach contend that metaphors are not special, but are rather at the core of ordinary language, indeed at the core of human experience. However, even if one rejects the notion of literal meanings, one must concede that there are a few basic concepts not understood metaphorically in order to avoid complete circularity. ‘There must be some grounding,’ as Lakoff and Turner (1989) put it,
rather metaphorically. Thus, such basic units as up-down, center-periphery, front-back (Lakoff, 1987, ‘The Spatialization of Form Hypothesis’ p. 283) must be conceived of as bases (i.e., ‘skeletal image schemas’ undergirding thought) upon which metaphors can stand. Without any grounding at all, it would be impossible to determine what is being characterized in terms of something else. If The sun is a jewel is no different from The sun is a star it makes no sense to speak of metaphor at all. That is, the very concept of metaphor implies that something is being said instead (metaphorically speaking, in place) of something else, something else for which it stands. Lakoff and Johnson (1999) explicitly claim that there is a vast system of literal concepts, for example, the basic-level concepts and spatial-relations concepts (p. 58), all of these grounded in sensorimotor experience. But according to Turner (1998), the difference between The sun is a star and The sun is a jewel is attributable not to different cognitive operations in the two cases, but to how conceptual connections between two mental arrays strike us, ‘depending on how those arrays are already related in our category structures. A connection seems literal or figurative (or somewhere in between) not absolutely but in relation to the category structures used to understand it’ (p. 62). But even if one grants that metaphors are pervasive in human language, their exact relation to thought is far from clear. As Gibbs (1998) has recently put it, ‘perhaps the most controversial claim about metaphor in the last 20 years is that this trope is not merely a figure of speech but a specific mental mapping that significantly influences how people think, reason, and imagine in everyday life’ (p. 89). In this paper we would like to look at an innovative way to address this claim, grounding our analyses in the principles of how the human brain functions.

3. Connectionism

Around the middle of the past century Hebb (1949) and Hayek (1952) proposed rather similar approaches to the mind–brain relation, Hayek emphasizing the mind aspect more than the brain aspect, with Hebb focusing more on the brain. Philosophically these approaches could be characterized as espousing a version neutral monism (see Bunge, 1980), i.e., that mind and brain are not of two substances, not of two orders, and that neither one is a ‘manifestation’ of the other; rather, the mental and the physical (neurological) order are one and the same, and represent two descriptions of the same entity. But since it is abundantly clear that neurons, synapses, and neurotransmitters are not the same sort of things as ideas, emotions, and perceptions, there must be some way to bridge this gap. The bridge, as explicitly argued by Hayek, is that the mental and physical orders have the same structure. A structure is an abstraction that can be instantiated in various media. The complex of lines, dots, and other markings written on pieces of paper when Beethoven composed the Eroica Symphony has the same structure as the piece when performed in the 21st century by the Cleveland Symphony. Similarly, human languages can be represented in speech, in writing, in manual signs, in Morse code, etc. The structure (ex hypothesis) is independent of the medium in which it is represented.
For Hayek (1952) the difference between the two orders (phenomenal and physical, i.e., neurological) is essentially between two vocabularies. The hierarchical nature of the nervous system is reflected exactly in the mental system (or vice versa). The principle by which the hierarchy is formed is called ‘classification’ or ‘categorization.’ Originally, impulses occurring repeatedly at the same time and place are classed together. These are then grouped with other similar classes. The categories are then classed with other categories, etc. The non-co-occurrence of two classes could also constitute a class. The nervous system progressively makes connections between categories and members of categories (the latter also being categories), thereby forming new classes. Some of these are established innately. Others occur in postnatal experience. The totality of all the connections and categorizations of categorizations of categorizations... looked at from the highest level of organization, constitutes the human person—the mind, when looked at from a mental perspective, the nervous system, looked at physically.

As an example of the application of this approach, Hayek discusses color perception. He claims that our phenomenal perception of color exactly parallels the physical mechanism of selective sensitivity to different wavelengths of light. His structural-realist solution is to say that the distinctions we make among different hues (‘hue’ being a phenomenal term) are just those distinctions made among wavelengths by the physical sensory and interpretive apparatus. (See Schnitzer, 1978, 1982, 1986 for further discussion of this approach as it relates to linguistic structure.)

Hebb (1949) is much more explicit regarding neuronal function and how structure can be built up from the inherent properties of neurons. A neuron is composed of a cell body (as is any cell), an axon, which carries electrical charge away from the cell body, and dendrites, whose membranes contain receptors for neurotransmitter chemicals. When a neuron fires, an electrical charge is sent down the length of the axon. At the end of the axon are terminals, which release neurotransmitter chemicals, some of which are excitatory, and some of which are inhibitory. The axon terminals are in close proximity with dendrites of a great many other neurons. The space between the axon terminal of one neuron and the dendrite of the other is called a synapse. The greater the concentration of neurotransmitter chemicals in the synapse, the more likely it is that charged ions will pass through the membrane openings of the dendrites of an adjacent neuron. The charge on the dendrite membrane is carried to the cell body, where it combines with the charge received by other dendrites of that neuron. Each neuron will have an activation threshold—a level which when reached or exceeded will cause it to fire. Since each dendrite is in close proximity with the axons of a great many neurons, whenever a neuron fires, it will be because the sum of the activation provided by the neurotransmitters of the various input neurons has added up to at least the minimum necessary to make the neuron fire. Since some neurons will be transmitting inhibitory neurotransmitter, it will be an algebraic sum. That is, if neuron N has a threshold of .6 (on scale of −1 to +1) to fire and if it receives excitatory transmission of .1, .3, .4, and .2 from four neurons A, B, C, and D, respectively, and inhibitory transmission of −.2, −.3, and −.2 respectively from three other neurons, X, Y, and Z, then neuron N will not fire because
.1 + .3 + .4 + .2 − .2 − .3 − .2 = .4, which is below the threshold of .6 necessary for neuron N to fire. Neuron N, when it does fire, may in turn cause neurons A, B, C, and/or D, to fire, either alone, or more commonly, in combination with other neurons whose axons interact with the dendrites of A, B, C, and/or D, either directly or by means of intermediate neurons, or both.

Whereas Hayek speaks of co-classification of impulses that occur repeatedly together, Hebb (1949) more specifically states, ‘When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased’ (p. 62). This is the process by which synaptic bonding between neurons is gradually strengthened or weakened. This statement, often known as ‘the Hebb Rule’ forms the basis of the connectionist approach we are advocating. Hebb goes further than Hayek, however, in explaining cognition neurologically by introducing his notion of the ‘cell assembly.’ According to Hebb’s model, groups of neurons that tend to fire together form a cell assembly whose activity can persist after the triggering event; the persistent cell assembly becomes the neuropsychological representation of that event. The history of each neuron’s activity will result in the links between it and others becoming stronger or weaker. Each neuron will have a particular threshold of activation necessary for it to fire. Depending on the strength of the links between it and other neurons, it will fire or not fire under various conditions.

In our illustrations of neural networks below, we will list strengths of links in terms of numbers between 0 and 1, in which 1 indicates a very strong link (i.e. the first neuron stimulates the second one very strongly) and 0 indicates no activation of the second neuron at all. That is, for purposes of illustration, we will consider only excitatory, not inhibitory, transmission.

4. Neural networks

With the advent of the modern computer, it became possible to create artificial functioning models along the lines of Hebb’s and Hayek’s proposals. These models, known as ‘connectionist,’ are intended to replicate brain structure at the appropriate level of analysis, i.e. the functional level; they are based on the concept of the ‘neural network’ (see Rummelhart and McClelland, 1986; McClelland and Rummelhart, 1986), and on the premise that how the mind works is how the brain works. Thus one makes brain models in order to understand the mind. In connectionism, each concept is represented by an interconnected set of ‘neurons’ that are connected to a great many other ‘neurons’ each giving different aspects of the concept. We place the word ‘neurons’ in scare quotes for two reasons: one is that in connectionist networks what we will call neurons are obviously not neurons in the biological sense; the claim is that they behave functionally in the way that biological neurons behave. The second is that what may be represented in a network as involving a single neuron, may be represented neurologically as groups of neurons or even whole arrays of them. For these reasons, in connectionist literature the term ‘unit’ (rather than
‘neuron’) is generally used. As McLeod et al. (1998) put it, ‘Contemporary connectionist models are based on the assumption that although they make simplifications they can provide a useful starting point for understanding how cognitive computations might be performed’ (p. 10). When we use the word ‘neuron’ henceforth, it is important to bear this caveat in mind. Another caveat is equally important: in the ensuing discussion, our highly abbreviated network proposals are intended merely as illustrative and pedagogical and in no sense represent actual proposals intended to account for the phenomena under discussion. The actual development of such networks is beyond the scope of this paper. It is hoped that this paper will stimulate research along these lines.

Let us try to understand the concept of neural network by way of example, by considering a network covering (a small subset of) the field of concepts of fruit. Table 1 is an illustration of what a tiny fragment of such a network might look like. This network, which would undoubtedly have thousands of neurons participating, and no doubt cover many more than five varieties of fruit, nevertheless can be illustrated by Table 1. The point is that each concept of fruit corresponds to a specific pattern of neuron firing. In the table, the top heading refers to each of the five neurons in the network. For each concept listed in the left-hand column, there is a notation of 0 or 1 for each of the five neurons. A 1 indicates that the corresponding neuron fires. A 0 indicates that the corresponding neuron does not fire. Hence the concept of a grape is represented by the pattern 11000, while that of a banana is represented by the pattern 10010. How the specific patterns would have arisen in the first place will depend on their history, along the lines of the Hebb Rule.

Let us now extend this notion by considering how one network could communicate with another. This is important if we want to consider how concepts relate to names, for example, or more generally, in Turner’s terminology, how conceptual patterns relate to formal ones.

What would make a banana concept be a banana concept in the first place would be the connections of its various neurons with other concepts, some in the same network, others in other networks. For example, banana would be connected with a network for CONCEPTS OF NATURAL COVERINGS and its covering would be connected with the word ‘peel’ in a network for NAMES OF NATURAL COVERINGS (including ‘skin,’ ‘bark,’ ‘rind,’ etc.). When the concepts peel and banana are both activated, their simultaneous activation would in turn activate the concept

<table>
<thead>
<tr>
<th>Neuron number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana concept</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Grape concept</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Apple concept</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pear concept</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Watermelon concept</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
yellow in a COLOR-CONCEPT network (and inhibit the concept green, for example) when the concept for ripe in some network of FLORA-stages-CONCEPTs is also activated; it would activate the concept green (and inhibit the concept yellow) when the concept for unripe is activated in the FLORA-stages network; the concept for blue would thus be inhibited by simultaneous activation of banana-concept and peel-concept no matter what FLORA stage is activated.

Let us now look at a specific illustration of how a concept from the FRUIT-CONCEPTS network could activate a name from a FRUIT-NAMES network. In connectionism, this will explain how we can know the name of a fruit when we think of the fruit. By similar means to those described below, we would know the name of a fruit when we see an example of it, by means of links between the FRUIT-NAME network and a FRUIT-IMAGE network. There would also be links between our FRUIT-CONCEPT network and the FRUIT-IMAGE network, which would allow us to associate our concepts of the various fruits with our images of them.

For purposes of exposition, let us focus on just two fruit concepts, banana and watermelon.

Let us first see how a connectionist account could associate the name ‘banana’ with the concept banana. The fragment of the FRUIT-NAME network will have only four neurons, for ease of exposition. As shown in Table 2, for bananas, the only FRUIT-CONCEPT neurons which fire are numbers 1 and 4. Thus 2, 3, and 5 do not participate in linking the concept banana to the FRUIT-NAME network. Hence only FRUIT-CONCEPT neurons 1 and 4 are referred to in Table 3.

The sum of the activation strengths for the two participating banana-concept neurons are listed in the third column. The fourth column lists the activation threshold for each of the four neurons in the FRUIT-NAME network. If the total equals or

Table 2
Fruit-concept network fragment for Banana and Watermelon showing strength of link to four neurons in fruit-name network

<table>
<thead>
<tr>
<th>Neuron number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana concept</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Watermelon concept</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Strength of link with neuron 1 of FRUIT-NAME Network</td>
<td>.1</td>
<td>.2</td>
<td>.3</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>Strength of link with neuron 2 of FRUIT-NAME Network</td>
<td>.3</td>
<td>.2</td>
<td>.1</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>Strength of link with neuron 3 of FRUIT-NAME Network</td>
<td>.1</td>
<td>.3</td>
<td>.1</td>
<td>.3</td>
<td>.2</td>
</tr>
<tr>
<td>Strength of link with neuron 4 of FRUIT-NAME Network</td>
<td>.2</td>
<td>.1</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
</tr>
</tbody>
</table>

Table 3
Links of the concept Banana to the FRUIT-NAME NETWORK

<table>
<thead>
<tr>
<th></th>
<th>BCN 1</th>
<th>BCN 4</th>
<th>Total</th>
<th>Threshold</th>
<th>Neuron fires?</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNN 1 link strength</td>
<td>.1</td>
<td>.4</td>
<td>.5</td>
<td>.4</td>
<td>1</td>
</tr>
<tr>
<td>FNN 2 link strength</td>
<td>.3</td>
<td>.1</td>
<td>.4</td>
<td>.5</td>
<td>0</td>
</tr>
<tr>
<td>FNN 3 link strength</td>
<td>.1</td>
<td>.4</td>
<td>.5</td>
<td>.5</td>
<td>1</td>
</tr>
<tr>
<td>FNN 4 link strength</td>
<td>.2</td>
<td>.2</td>
<td>.4</td>
<td>.6</td>
<td>0</td>
</tr>
</tbody>
</table>

BCN = BANANA-CONCEPT neuron; FNN = FRUIT-NAME neuron.
exceeds the threshold, the corresponding neuron fires. Thus the word ‘banana’ is represented in the FRUIT-NAME network by the pattern 1010, indicating that FRUIT-NAME neurons 1 and 3 fire and that 2 and 4 do not.

For comparison let us now look at Table 4, which presents the fragment of the FRUIT-NAME network for ‘watermelon.’

For watermelons, the FRUIT-CONCEPT neurons which fire are numbers 1, 3, and 5 (see Table 2). Thus 2 and 4 do not participate in linking the concept watermelon to the FRUIT-NAME network.

This table indicates that the word ‘watermelon’ is represented in the FRUIT-NAME network by the pattern 1101, indicating that FRUIT-NAME neurons 1, 2, and 4 fire and that 3 does not.

We now return to the question of how the various neuronal patterns and thresholds for neuron firing would have arisen in the first place. Recall that the Hebb Rule states that when one neuron fires upon being stimulated by another neuron repeatedly, it increases the probability that it will fire, that is, it lowers its threshold. One can conceive of a neonate brain (and mutatis mutandis, a neonate connectionist network) as having all the neurons connected with one another (either directly or by means of connections through intermediate neurons) without any preferential connections (i.e., all connections are equally weak). In the case of a real organism, of course, many preferential connections are innately ‘hard wired.’ (And in modern connectionist modeling of cognitive processes, an initial bias is often programmed into the network.) Eventually, learning what counts as an apple and what counts as a pear, for example, is done by means of those neurons in visual cortex that produce images of apples making the firing of the specific pattern of neurons in the FRUIT-NAME network and in the FRUIT-CONCEPT network become more ‘efficient,’ in Hebb’s words—i.e., lowering their threshold for firing. What might start out as an undifferentiated apple–pear concept would become differentiated as the neuronal patterns became established for the two words ‘apple’ and ‘pear,’ and for their connections to the two concepts apple and pear.

5. Relevance to metaphor

In a persuasive essay, Turner (1998) characterizes the notion of figure as ‘pairing between formal and conceptual patterns’ (p. 60). He rejects the distinction between

---

Table 4
Links of the concept Watermelon to the FRUIT-NAME NETWORK

<table>
<thead>
<tr>
<th>WCN 1</th>
<th>WCN 3</th>
<th>WCN 5</th>
<th>Total</th>
<th>Threshold</th>
<th>Neuron fires?</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNN 1 link strength (from Table 2)</td>
<td>.1</td>
<td>.3</td>
<td>.2</td>
<td>.6</td>
<td>.4</td>
</tr>
<tr>
<td>FNN 2 link strength</td>
<td>.3</td>
<td>.1</td>
<td>.1</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>FNN 3 link strength</td>
<td>.1</td>
<td>.1</td>
<td>.2</td>
<td>.4</td>
<td>.5</td>
</tr>
<tr>
<td>FNN 4 link strength</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.6</td>
<td>.6</td>
</tr>
</tbody>
</table>

WCN = BANANA-CONCEPT neuron; FNN = FRUIT-NAME neuron.
‘literal’ and ‘figurative’ meaning as a ‘psychological illusion’ (p. 60) based on a folk theory. ‘In this folk theory, a connection is true if the state of affairs to which it refers is the case in the world (i.e., it is the case [for the example Big trees grow slowly] that everything that is both big and a tree has the complex feature grows slowly). It is false if the state of affairs to which it refers is not the case in the real world’ (p. 61). For Turner, the difference between The sun is a star and The sun is a jewel is thus not attributable to different cognitive operations but rather to how conceptual connections between two mental arrays strike us, ‘depending on how those arrays are already related in our category structures. A connection seems literal or figurative (or somewhere in between) not absolutely but in relation to the category structures used to understand it. ‘A child is a light bulb’ asks us to connect mental arrays that are basic level categories, and thus seem figurative (p. 63). Also related is degree of ‘generative entrenchment.’ Thus, in I am making intellectual progress, a thinker is not ‘literally’ a traveler, but this idea is so entrenched that it is not thought of as figurative.

In our approach, we define an explicit metaphor as one of the form ‘X is (a) Y’. We interpret such a metaphor as an explicit instruction to link one connectionist network with another. To the extent that the networks are already linked, the metaphor lacks force: it is no longer perceived as a metaphor. Thus Life is a journey and its corollaries does not strike us as particularly metaphorical, given the strength of the links between the networks containing aspects of life and journeys (with obvious links between the two networks involving leaps, setbacks, progress, stagnation, side-tracks, etc.).

But metaphors are most often not expressed explicitly in the form ‘X is (a) Y’. Consider the example ‘ethnic cleansing.’ In this example, although there is no explicit instruction to link one network to another, the use of two words in a modifier-head construction forces the perceiver (or creative first-time user) to deal with a construction containing components drawn from different networks, ones that are not usually linked (since cleansing is usually linked with physical objects) and ‘ethnic’ would not be found in such a network. The myriad networks in which ‘ethnic’ would be found would not be linked to a network containing ‘cleansing.’ But we would interpret the use of the term ‘ethnic cleansing’ as an implicit instruction to link a network in which ‘ethnic’ occurs with one in which ‘cleansing’ occurs. In Turner (1998) terminology, the two arrays are not already related in our category structures. Thus, the instruction to relate them makes the combination seem metaphorical. As Turner notes, the combination ‘ethnic cleansing’ ‘was judged to be highly figurative when first used, but the effect seems to be wearing off with frequent exposure’ (p. 63).

Why is the effect wearing off? In connectionist terms, the link between ‘ethnic’ and ‘cleansing’ is strengthened with each hearing or use of the phrase. This follows directly from the Hebb Rule.

But why should metaphors exist in the first place? In connectionist theory, learning takes place by the gradual strengthening of synaptic connections through repetitions of neurons’ firing causing other neurons to fire (as well as the gradual weakening of connections through no longer repetitively firing, or by inhibition).
Knowledge consists of patterns of neuronal activity which are modified as we learn. Because learning takes place incrementally, it is easier to acquire knowledge that requires minimal modification of the neuronal connection patterns than knowledge that requires major reconnection (Goldblum, 2001). This is why it is easier to learn new material that is just incrementally more advanced than what is already known. This phenomenon has been researched and discussed extensively in the field of second-language acquisition. According to Krashen’s (1985) ‘input hypothesis,’ in order for second language acquisition to take place, ‘we progress... by understanding input that contains structures at our next ‘stage’—structures that are a bit beyond our current level of competence. (We move from \( i \), our current level, to \( i + 1 \), the next level along the natural order, by understanding input containing \( i + 1 \).)’ (p. 2).

Although Krashen does not cite him, this position is similar to that of Vygotsky, who, discussing learning in general, rather than language acquisition, defined a ‘zone of proximal development’ as ‘the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers’ (Vygotsky, 1978, p. 86; the original appeared in Russian in 1935). Goldblum (2001) relates the connectionist approach to learning to Vygotsky’s ‘zone of proximal development.’ Also see Lantolf and Appel (1994) for more on ‘Vygotskyan’ approaches to language acquisition.

The point of all this is that incremental learning—learning by making use of what is already known—is easier, more efficient, than large-scale learning of new knowledge. The instruction to link one network with another made explicitly or implicitly by a metaphor is a way of allowing one to increase knowledge by making use of what is already known. Linking ‘ethnic’ with ‘cleansing’ allows both terms to be understood in terms of connections already in place but not linked to each other.

6. Why connectionism?

What makes the connectionist approach unique is its grounding in the structure of the nervous system. The approaches to metaphor that one finds in the literature, whether of the classical or the cognitivist kind, are all similar in that they do not concern themselves with how metaphor is represented neurophysiologically. A tenet of many

---

2 Turner (1991) discusses the neurophysiological basis for conceptual connections from the perspective of various kinds of ‘conceptual patterns’ that are activated and deactivated, referring to the most frequently activated conceptual connections as ‘use-entrenched’ (part of his explanation for how metaphors become conventional), and even ‘generatively entrenched’ if other conceptual patterns frequently ‘depend’ upon them. Most significantly, Turner sees the mind’s conceptual activities as a process, dynamic and plastic, not static: ‘A conceptual system consists of dynamic and plastic patterns that are lost if not used. Concepts themselves are not fixed objects but rather activities; they do not have essences but rather functions. The mind is not a machine that works on objects, it is rather a process, a dynamic process, and it is plastic’ (pp. 156–157).
approaches to understanding the mind–body relation is that the key issue is considered to be what it is that is represented, rather than the structure of the representation itself. Proponents of such approaches may be called ‘representationalists’ (cf. Goldblum, 2001). One of the leading spokesmen of this approach, Fodor (1975, 1981, 1983, 1987, 1990) has been outspoken in claiming that the human mind is a system of modules whose representation is not necessarily constrained by specific neurophysiological hardware. Proponents of this view (including Noam Chomsky, Stephen Pinker, Zenon Pylyshyn, and many other influential scholars in linguistics and psychology) believe that the brain is sufficiently plastic as to be able to use a wide variety of hardware vehicles to ground a given representation. These researchers broadly conceive of the brain as functioning like a modern digital computer, capable of running a wide variety of software with a wide variety of hardware specification. (See Newmeyer, 2003, for arguments against connectionist approaches to grammar.) For them, what is important is what is represented; questions of how can be dealt with later by experts in brain hardware. In contrast, connectionism rejects that hardware–software distinction when it comes to brain function. Each change in activation threshold is a change both in hardware and software—there is a change in how the neurons are connected with one another.

But if one’s theory is not directly grounded in the way that nervous systems works, then arguments for one proposal over another can only resort to evidence from behavioral data. But these data will grossly underdetermine the choice of proposal: in general, there will be many possible proposals consistent with such data. In contrast, our structural-realist proposal requires that proposals be couched in terms of neurophysiological processing, regardless of how far abstracted from real neuronal function.

Because of the fact that both connectionist and nonconnectionist approaches make use of representations, we prefer the dichotomy vehicle-process to connectionism–representationalism to refer to what is at issue. O’Brien and Opie (1999) discuss the issue of phenomenal experience from the perspective of the computational theory of mind, ‘the theory that treats human cognitive processes as disciplined operations defined over neurally realized representations’ (p. 128). If the nature of the neurological representational vehicles is not considered primary, if the focus is on what they do rather than what they are, then a theory may be considered to be a process theory. On the other hand, if the represented phenomena are identical (as in a structural-realist conception) to the vehicles of representation, then a theory may be considered to be a vehicular theory (O’Brien and Opie, 1999). What we wish to emphasize at this point is that all of the approaches to metaphor we have thus far mentioned have

---

3 We use the term vehicular rather than vehicle to avoid confusion with the use in describing metaphors of the word vehicle in the sense that Richards (1936) coined to replace ‘clumsy descriptive phrases’ as ‘the borrowed [idea]’, ‘what [what is really being said] is compared to’, ‘the imagined nature’, ‘what [the principal subject] resembles’, or the ‘image [of the idea]’ (p. 96). The vehicle in Richards’ sense is now referred to by several other terms including the source, the base, the second term, or simply the figure.
been implicitly or explicitly couched in process theory. None of them, even those that speak of ‘mental arrays’ or ‘base and target domains’ are vehicular in this sense. Thus none of them makes any claim as to cerebral function.

By couching our analysis in our vehicular approach we will show why some of the disputes are moot, that is to say that apparently different claims turn out not to have different content when grounded in a connectionist account. For example, as noted above, Gibbs (1998) has recently claimed that metaphor is ‘a specific mental mapping that significantly influences how people think, reason, and imagine in everyday life.’ In addressing this claim, Gibbs offers four ‘psychological hypotheses’ that ‘reflect a hierarchy of possibilities about the interaction between metaphoric patterns of thought and different aspects of language use and understanding’ (p. 93). We would like to consider Gibbs’s hypotheses and show how a connectionist approach could clarify these positions.

**Hypothesis 1.** Metaphoric thought plays some role in the historical evolution of what words and expressions mean (p. 93).

That metaphoric thought plays some role in the historical evolution of what words and expressions mean could hardly be doubted by anyone who has ever consulted an etymological dictionary. Consider the entry for the Indo-European root of ‘gen-’ (to give birth, to beget) in the American Heritage Dictionary, Fourth Edition, Houghton-Mifflin, 2000, New York, p. 2028). All of the following English words derive from this root, nearly all of them metaphorically: gender, general, generate, generation, generic, generous, genre, congener, degenerate, engender, miscegenation, gene, genus, genealogy, genocide, genotype, heterogeneous, genial, genius, congenial, engine, ingenious, indigenous, genuine, ingenuous, germ, German, germane germinal, germinate, gonad, king, kind, genteel, gentle, gentry, jaunty, gendarme, genesis, kindergarden, genial, genitive, genitor, gingerly, congenital, primogenitor, primogeniture, progenitor, progeny, benign, malign, pregnant, impregnate, naïve, nascent, natal, nation, native, nature, agnate, cognate, connate, innate, neonate, puny, renaissance, as well as other, less familiar words.

Whatever may have happened historically, this evidence says relatively little about real-time processing of metaphor.

From a connectionist perspective, what these examples may indicate is that at some time in the past, links between different networks were formed. At the time, these connections were innovative and involved multiple networks. Over time the connections have become so well established as to become integrated into the relevant networks themselves, losing first the sense of being metaphorical, and finally even the sense of being related. Certainly heterogeneous and benign are unlikely to be linked in anyone’s mind–brain. Their links have become weakened to the point of disappearing.

**Hypothesis 2.** Metaphoric thought motivates the linguistic meanings that have currency within linguistic communities, or may have some role in speakers’/hearers’ presumed understanding of language (p. 94).
Examples cited by Gibbs from Lakoff and Johnson’s (1980) metaphorical system of understanding LOVE IS A JOURNEY include, Look how far we’ve come, It’s been a long, bumpy road, We’re at a crossroads, and We may have to go our separate ways. According to Gibbs, the reason why these expressions cluster the way they do is that our conception of love is metaphorically structured. The arguments here are persuasive (see Lakoff and Johnson, 1980). But from a connectionist perspective, the fact that people do not generally think of these expressions as being metaphorical may indicate that the connections originally made between a network containing the concept of love (with perhaps concepts of other emotions in the same network) and a network containing the concept of JOURNEY, have become so well established as to lose their sense of novelty. The emphasis in this hypothesis is on ‘currency within linguistic communities.’ Although they may continue to structure our thought, as metaphors they are dead.

**Hypothesis 3.** Metaphoric thought motivates real-life, contemporary speakers’ use and understanding of why various words and expressions mean what they do (p. 96).

Gibbs distinguishes this hypothesis from the second one on the basis of evidence from psycholinguistic experimentation. In investigating the metaphorical structure of ANGER IS HEATED FLUID IN A CONTAINER (Kovecses, 1986; Lakoff, 1987), resulting in such expressions as blow your stack, flip your lid, and hit the ceiling (p. 98), he found that when subjects understand such anger idioms as these, they infer that ‘the cause of anger is internal pressure, that the expression of anger is unintentional, and that it is done in an abrupt, violent manner,’ but that they ‘do not draw the same inferences about causation, intentionality, and manner when comprehending literal paraphrases of idioms, such as “get very angry” ’ (p. 99). A connectionist interpretation would here posit a more currently active connectivity between an emotion-concept network and a network containing the linked concepts of contained heated fluid. Subjects understand these anger idioms in terms of activation of this latter network by the anger concept in the emotion-concept network.

**Hypothesis 4.** Metaphoric thought functions in people’s immediate on-line use and understanding of linguistic meaning (p. 103).

As Gibb’s states, although many would be willing to accept hypothesis 1, 2, or 3, in which preexisting metaphorical constructs may influence how people understand

---

4 Turner (1991), drawing on the earlier work of Lakoff (1987) sees this phenomenon of metaphors becoming conventional over time and thereby losing their novelty as a significant effect of the ‘invariance hypothesis’: ‘From this perspective, metaphoric understanding is an extremely imaginative process, and basic metaphors, far from being rigid little automatic preformed packets that save us from having to use our imaginations, are rather cases where imaginative connections have become conventional and where mappings that are ontologically constituted by imaginative links have become, through reinforcement, entrenched’ (p. 180).
idiomatic meaning, many reject the idea of immediate direct metaphoric comprehension. Gibbs cites recent studies that indicate that metaphors ‘influence people’s judgments of the appropriateness of idioms,’ but are not accessed during ‘immediate idiom comprehension’ (p. 104), as measured in reading-time tasks (Glucksberg et al., 1993). However, we would suggest that fresh, perhaps newly coined metaphors would have to be processed in this way to be understood. Metaphors of this type would represent (explicit or implicit) instructions to make connections with some other network. Perhaps the negative evidence came from relying on metaphors that were already too well known to the subjects and that therefore already displayed strong connections between the terms of the tenor and the terms of the vehicle.

7. Conclusion

From our standpoint, Gibbs’s four hypotheses are not four possibilities about how metaphor might be mentally encoded. Rather, they represent four stages along a connectionist continuum. Ranging from the connection so ancient as to be no longer part of a network to which it was strongly connected at one point, through various stages of prior connections representing various strength of connections among networks, to the fresh metaphor in which the network of the tenor is connected for the first time to the network of the vehicle.

Clearly, what remains to be done is to implement connectionist theory in the construction of actual networks relevant to the issue of metaphor. This should be done in two ways. First of all, networks should be constructed which illustrate the difference between the type of metaphor that is already entrenched and the kind that strikes one as innovative and all the levels in between. But more importantly, to really confirm our thesis, there remains the task of constructing connectionist networks capable of actually inventing new metaphors; that is, given a certain input pattern, a connectionist network should derive an output pattern that is acceptable to human observers as a novel metaphoric interpretation of the input pattern.

Work has already begun in this regard. For example, Lakoff and Johnson (1999) cite doctoral work by Narayanan (1997) who extended neural modeling of motor schemas to a model of conceptual metaphor. He considered body-based metaphors, such as ‘France fell into a recession and Germany pulled it out.’ He found that ‘models of the motor schemas for physical actions can—under metaphoric projection—perform the appropriate abstract inferences about international economics’ (Lakoff and Johnson, 1999, p. 42). But as Lakoff and Johnson readily admit, results such as these show only that ‘abstract reasoning using those schemas can be carried out using neural motor-control simulations’ (p. 42, italics ours). They do not show that people actually use perceptual and motor-control brain networks to perform such reasoning.

Gibbs (1994) and Lakoff and Johnson (1999) both cite some unpublished studies of Allbritton (1992) which indicate that subjects are quicker at recognizing sentences from a previously presented text when the subjects are primed by a sentence that is part of the same extended conceptual metaphor as the sentence being considered than they are when they are primed by a sentence not part of that metaphor. Thus,
subjects recognized the sentence ‘Sparks flew the moment they first saw one another’ more rapidly when primed by the sentence ‘The attraction between John and Martha was overwhelming,’ which is part of the LOVE IS A PHYSICAL FORCE metaphor, than when primed by the sentence ‘John and Martha met at a party a month ago,’ which is not. These results are consistent with our connectionist proposal: Assuming that there is a physical-force network of some kind, reaction time should be shorter for tasks that require staying within the network than for those requiring a connection to the network from outside of it.

Recently, Bryan Meier and colleagues (Meier et al., 2004; Meier and Robinson, 2004) have presented results of experiments that suggest that in fact at least some kinds of metaphoric expression are obligatory, that is the connection between the vehicle and tenor is not subject to the subject’s discretion, but rather that metaphoric thought does function in immediate on-line use and understanding of linguistic meaning. In the first paper, Meier et al presented 50 words with ‘positive meaning’ (such as clean, gentle, hero, trust, cordially) and 50 words with negative meanings (such as cruel, steal, ugly, divorce, enemy) to subjects who were asked to rate the words as positive or negative. The words were randomly presented in either white or black letters. Subjects evaluated positive words more quickly and more correctly when they were presented in white letters, whereas they evaluated negative words more quickly and more correctly when they were presented in black letters. The basis for this study is the primary or basic metaphor LIGHT IS GOOD. In these studies the color of the stimuli affected the subjects’ responses in spite of the fact that it was completely irrelevant to the task of evaluating the words as positive or negative.

In the second paper, Meier and Robinson probed the basic metaphor UP IS GOOD. In this study, the same sets of positive and negative words were presented randomly projected to the upper or the lower portion of a screen. Predictably, the subjects evaluated positive words more quickly and more correctly when they were projected on the upper half of the screen, whereas they evaluated negative words more quickly and more correctly when they were projected on the lower half.

In both sets of studies, the fact that subjects apparently were not able to exert independent control over the tenor without interference from the vehicle in the metaphorical relations being tested indicates that there is some automatic—involuntary—connection between neuropsychological representations of the source and the target, along the lines proposed by the proponents of the congnitivist approach to metaphor. These results provide strong support for Gibbs’s fourth hypothesis—Metaphoric thought functions in people’s immediate on-line use and understanding of linguistic meaning—and they are consistent with the connectionist neural-network approach we have been advocating: In these cases the networks are very strongly connected indeed.

References


