PERGAMON

Language & Communication 22 (2002) 313–329

LANGUAGE & COMMUNICATION

www.elsevier.com/locate/langcom

Radical connectionism: thinking with (not in) language

Gerard O'Brien*, Jon Opie

Department of Philosophy, University of Adelaide, South Australia 5005, Australia

Abstract

In this paper we defend a position we call radical connectionism. Radical connectionism claims that cognition *never* implicates an internal symbolic medium, not even when natural language plays a part in our thought processes. On the face of it, such a position renders the human capacity for abstract thought quite mysterious. However, we argue that connectionism is committed to an analog conception of neural computation, and that representation of the abstract is no more problematic for a system of analog vehicles than for a symbol system. Natural language is therefore not required as a representational medium for abstract thought. Since natural language is arguably not a representational medium *at all*, but a conventionally governed scheme of communicative signals, we suggest that the role of internalised (i.e. self-directed) language is best conceived in terms of the coordination and control of cognitive activities within the brain.

Keywords: Analog; Computation; Connectionism; Representation; Resemblance; Thought

1. Introduction

It is undeniable that the cognitive divide between ourselves and other animals is intimately connected with our capacity to comprehend and produce natural language. But exactly what this connection consists in is a matter of some controversy. Is natural language the *basis* of the divide or merely a *consequence* of it? That is, does the ability to deploy a natural language enable a form of cognition that is unavailable to infra-verbal animals, or is that ability a result of the difference between cognition in humans and other animals?

The *classical* computational theory of mind—which holds that cognition is the disciplined manipulation of symbols in an innate language of thought—opts for the

* Corresponding author. Tel.: +61-8-8303-4455.
E-mail addresses: gerard.obrien@adelaide.edu.au (G. O'Brien), jon.opie@adelaide.edu.au (J. Opie).

0271-5309/02/$ - see front matter
PII: S0271-5309(02)00010-1

latter response.[1] According to this position, all thought, no matter where it occurs in the animal world, is carried out in a linguiform representational medium, and hence the evolution of *natural* language didn't mark the development of a novel form of cognition. Instead, that evolution is itself to be (somehow) explained in terms of augmentations to the underlying functional architecture of the human brain—augmentations that account, first and foremost, for our enhanced cognitive capacities. From the classical perspective, therefore, natural language is a by-product of the representational medium of human thought, rather than partly constitutive of it.

The view from connectionism, the now popular alternative to classicism in cognitive science, is more complicated.[2] Connectionist networks don't compute by manipulating symbols, and hence don't deploy a linguiform representational medium. As a consequence, connectionists can regard the role of natural language in human cognition in two very different ways.

The first way, which we might call *ecumenical* connectionism, holds that the evolution of natural language resulted in a novel form of cognition, since it enabled connectionist networks to implement classical-style computation. On this view, the cognitive divide between ourselves and other animals is indeed a *computational* one. Even though much of human cognition (especially perceptual cognition) implicates a non-symbolic representational medium, rendering it continuous with cognition in other animals, our brains somehow bootstrap their way to genuine symbol-processing by way of natural language, and are thus in some respects computationally unique. We do, at least in part, think in natural language. Moreover, doing so enables us to enter realms of abstract thought that are inaccessible by any other means.[3]

The second way, *radical* connectionism, rejects this hybridisation. It shares with classicism the view that all of human cognition, including our capacity to deploy a natural language, depends on computational resources much like those that underpin the cognitive achievements of infra-verbal animals. But radical connectionism differs from classicism in that it rejects any role for a linguiform representational medium. Not only don't we think in our natural language, we don't think in language whatsoever.

On the face of it, radical connectionism would seem to be at a disadvantage with respect to both classicism and ecumenical connectionism. For how is it possible to engage in abstract thought without exploiting a symbolic representational medium? It is for this reason, perhaps, that radical connectionism is under-subscribed in the

[1] At least, this is the view from the standard formulation of classicism, as developed most notably by Fodor (1975, 1987). There are a few classicists who hold that the language of thought is the subject's natural language rather than an innate "mentalise" (see, e.g. Harman, 1973, and Devitt and Sterelny, 1987). Such theorists thus share with ecumenical connectionists (see later) the view that humans think, at least in part, in natural language.

[2] For useful introductions to connectionism, see Bechtel and Abrahamsen (1991), Clark (1989 Chaps. 5–6) and Tienson (1987).

[3] Perhaps the first place such an ecumenical version of connectionism was outlined was Rumelhart et al. (1986). Since that time the position has been defended, for example, by Smolensky (1988), Bechtel and Abrahamsen (1991, Chap. 7) and, most comprehensively, Clark (1989, Chap. 7; 1997, Chap.10).

literature. While there are plenty of connectionists willing to bet that large parts of human cognition are achieved without symbolic representational resources, there are very few who think that all of it can be.[4] The main debate here, it would seem, is between classicists and ecumenical connectionists, and turns on the question whether we think our abstract thoughts in mentalese or natural language.

Despite this disadvantage, in this paper we seek to defend a version of radical connectionism. Our proposal has two key elements. The first is a story about the nature of the representing vehicles that connectionist networks deploy. We'll argue that although these vehicles are non-symbolic, their representational content can be highly abstract. The second is a claim about the catalysing role that natural language plays in higher cognition. We'll argue that while natural language doesn't constitute the representational medium of abstract thought, it nonetheless facilitates such thought by supplying a system of communicative signals which coordinates and controls the cognitive activities of connectionist networks in far flung regions of the brain. The proposal, in other words, is that we think *with* language, rather than *in* it.

2. A structural resemblance theory of connectionist representation

Human cognitive processes, according to connectionism, are the computational operations of a multitude of connectionist networks implemented in the neural hardware in our heads.[5] Our aim in this section is to outline a theory of representation which indicates how connectionist representing vehicles, despite being non-symbolic, are capable of highly abstract representational contents. In order to do this, however, we need to know a little about the representing vehicles that connectionist networks employ. This is where we begin.

2.1. Connectionist representing vehicles

A connectionist network is a collection of interconnected processing units, each of which has an *activation level* that is communicated to the rest of the network via modifiable, weighted connection lines. From moment to moment, each unit sums the weighted activation it receives, and generates a new activation level that is some threshold function of its current activity and that sum. A connectionist network typically performs computational operations by "relaxing" into a stable *pattern of activation* across its constituent units, in response to the input it receives. This relaxation process is mediated by the connection weights, because they determine how, and to what extent, activation is passed from unit to unit.

[4] Indeed, the only theorist we know of who comes close to defending radical connectionism is Paul Churchland— see, e.g. Churchland (1995, pp. 257–264; in preparation, Sec .8).

[5] In this context connectionist networks are to be understood as *idealised* models of real neural networks, which, although unrealistic in certain respects, capture the computationally significant properties of neural networks (see, e.g. Churchland and Sejnowski, 1992, Chap.3; O'Brien, 1998; Opie, 1998)

The representational capacities of connectionist networks rely on the plasticity of the connection weights between the constituent processing units.[6] By altering these connection weights, one alters the activation patterns the network produces in response to its inputs. As a consequence, an individual network can be taught to generate a range of stable target patterns in response to a range of inputs. These stable patterns of activation, because they are generated rapidly in response to the flux of input impinging on individual networks, constitute a transient form of information coding, which we will refer to as *activation pattern representation*.

In connectionist theorising, activation patterns are the entities that receive an interpretation, such that each pattern of activation across a network has a distinct semantic value (often specified in terms of a semantic metric). In this respect activation pattern representations are akin to the tokens on the tape of a Turing machine.[7] An individual pattern, just like a symbol on the tape, is an element in a system of physically structured objects for which there is a semantics (a mapping between individual representing vehicles and some represented domain), and a "parser" mechanism that is capable of recognising and responding to semantically significant variations in physical structure. In the case of a Turing machine the parser is the read/write head through which the tape passes. An activation pattern is "parsed" by virtue of having effects on other networks. Given this, we believe it is warranted to apply to connectionism the now standard terminology and say that stable activation patterns represent information in an *explicit* fashion.[8]

While activation patterns are a transient feature of connectionist networks, a "trained" network has a relatively long-term capacity to generate a set of distinct activation patterns, in response to cueing inputs. So a network, in virtue of its connection weights, can be said to *store* appropriate responses to input. This form of information coding, which is sometimes referred to as *connection weight representation,* is the basis of long-term memory in connectionist systems. Such long-term storage of information is superpositional in nature, since each connection weight contributes to the storage of every stable activation pattern that the network is capable of generating. Consequently, the information that is stored in a network is not encoded in a physically discrete manner. The one appropriately configured network encodes a *set* of contents corresponding to the set of activation patterns it is capable of generating. Such contents are not explicit; they are merely *potentially explicit* (Dennett, 1982, p.216–217).

[6] For good general introductions to the representational properties of connectionist systems, see Bechtel and Abrahamsen (1991, Chap. 2), Churchland (1995), Churchland and Sejnowski (1992, Chap. 4) and Rumelhart and McClelland (1986, Chaps. 1–3).

[7] Though, as we noted in the first section, one important respect in which activation pattern representations differ from classical symbols is that their semantics is not language-like. Symbol structures, unlike activation pattern representations, have a (concatenative) combinatorial syntax and semantics. The precise nature of the internal structure of connectionist representations is a matter of some debate (see, e.g. Fodor and Pylyshyn, 1988; Smolensky, 1987; van Gelder, 1990)

[8] For a detailed argument to this effect see O'Brien and Opie (1999, pp. 133–137); for discussion see Clapin and O'Brien (1998).

Potentially explicit information is encoded in a connectionist network in virtue of its relatively long-term capacity to generate a range of explicit representations in response to cueing inputs. This capacity is governed by a network's configuration of connection weights. However, since a network's connection weights are also responsible for the manner in which it responds to input (by generating activation pattern representations), this means that the mechanism driving the computational operations of a connectionist network is identical to the mechanism responsible for its long-term storage of information. So there is a strong sense in which it is the potentially explicit information encoded in a network (the network's "memory") that actually governs its computational operations. This fact has major consequences for the connectionist take on cognitive processes. Crucially, information that is merely potentially explicit in connectionist networks need not be rendered explicit in order to be causally efficacious. The information that is encoded in a network in a potentially explicit fashion is causally active whenever that network responds to input. With this very brief account of connectionist representing vehicles before us, it is now time to consider how these vehicles acquire their representational content.

2.2. Computational architecture and representational content

The task of a theory of representational content is to explain how nervous systems can be in the representing business in the first place—how brain states can be about aspects of the world. It is a commonplace in the philosophy of mind that a theory of representational content must be *naturalistic*, in the sense that it cannot appeal to properties that are either non-physical or antecedently representational.[9] Given this constraint, there would seem to be just two different objective relations that the brain's representing vehicles are capable of bearing to the world, and which might therefore form the basis of a naturalistic account of representational content. These are *causation* and *resemblance*.[10] Which of these relations is the most appropriate, in our view, is determined by the brain's computational architecture.

Classicism operates with a *digital* conception of neural computation, and a *symbolic* conception of the brain's representing vehicles. The computational capacities of a digital device are embodied in the rules that regulate the behaviour of its explicit representing vehicles, rather than in the structural properties of the vehicles themselves (Fodor, 1987). Classicism has thus fostered a climate (in both cognitive science and the philosophy of mind) in which the theoretical focus is directed mainly at the computational/causal *relations* that representing vehicles enter into, and not at their *intrinsic* properties. This theoretical focus has had a significant impact on the development of theories of representational content. On the one hand, it has completely inhibited the development of resemblance approaches to representational

[9] See, e.g. Cummins (1989, pp.127–129; 1996, pp. 3–4), Dretske (1981, p.xi), Field (1978, p.78), Fodor (1987, pp. 97–98), Lloyd (1989, pp. 19–20), Millikan (1984, p. 87) and Von Eckardt (1993, pp.234–239).

[10] See, e.g. Von Eckardt (1993, pp.149–152). Some philosophers think that convention is a third possibility, but this is controversial, since it is not clear that convention is consistent with the naturalism constraint.

content, since, as Cummins observes, "[classical] computationalists must dismiss similarity theories of representation out of hand; nothing is more obvious than that [symbolic] data structures don't resemble what they represent" (Cummins, 1989, pp. 30–31). And on the other, it has encouraged the development of causal theories of content, since causation would appear to be the one objective relation that symbols are capable of bearing to the world.

The computational capacities of a connectionist system, by contrast, are not inherited from rules that are distinct from the intrinsic properties of its representing vehicles. Indeed, as we saw in the previous subsection, connectionism dispenses with the classical distinction between representing vehicles and the processes that act on them (the so-called code/process divide—see Clark, 1993). The substrate that stores, in potentially explicit form, everything that a network "knows" (i.e. the network's configuration of weighted connections) is the very mechanism that governs its computational operations. Connectionist devices achieve their computational competences not by applying rules to the representing vehicles they generate, but by deploying learning procedures which gradually shape these vehicles so that they come to resemble aspects of the task domains over which they operate (O'Brien, 1999).

Consider, as an example, NETtalk (Sejnowski and Rosenberg, 1987). NETtalk transforms English graphemes into contextually appropriate phonemes. This task domain is quite abstract, comprising the letter-to-sound correspondences permitted in the English language. Back-propagation is used to shape NETtalk's activation landscape—which comprises all the potential patterns of activity across its 80 hidden units—until the network performs accurately. Once it is trained up in this fashion, there is a systematic relationship between the network's activity and the target domain, such that variations in activation patterns systematically mirror variations in letter-to-sound correspondences. It is this resemblance relation that is revealed in the cluster analysis which Sejnowski and Rosenberg applied to NETtalk. And it is this resemblance relation that makes it right and proper to talk, as everyone does, of NETtalk's having a *semantic metric*, such that its activation landscape becomes a *representational* landscape.

When such a resemblance relation exists between a network's representing vehicles and its task domain, there is no need to apply rules to those representing vehicles in order to govern their processing. Instead, the computational processes of the network are governed by the model of the task domain that it embodies. Thus, when NETtalk is exposed to an array of graphemes, the resemblance relation embodied in its connection weights automatically produces the contextually appropriate phonemic output.

The upshot of all of this is that the computational capacities of a connectionist system are embodied in the intrinsic properties of its representing vehicles (see Section 2.4). As a consequence, the almost exclusive focus in contemporary philosophy of mind on the *causal relations* into which the brain's representing vehicles enter is no longer wholly appropriate. Connectionism brings with it an additional focus on the *intrinsic properties* of the representing vehicles themselves. Indeed, connectionism compels us to explain representational content in terms of resemblance relations

between the intrinsic properties of the brain's representing vehicles and their target domains.[11] For this reason connectionist representations are not symbols, but *analogs* — representing vehicles whose physical form bears a non-arbitrary relationship to the objects they represent.

But exactly what kind of resemblance relations are required to ground the representational content of connectionist representing vehicles? We take up this question in Section 2.4. Before doing so we consider resemblance more generally.

2.3. Varieties of resemblance

Resemblance is a fairly unconstrained relationship, because objects or systems of objects can resemble each other in a huge variety of ways, and to various different degrees. However, one might hope to make some progress by starting with simple cases of resemblance, examining their possible significance for connectionist representing vehicles, and then turning to more complex cases. Let us begin, then, with resemblance between concrete objects. The most straightforward kind of resemblance in this case involves the sharing of one or more physical properties. Thus, two objects might be of the same colour, or mass, have the same length, the same density, the same electric charge, or they might be equal along a number of physical dimensions simultaneously. We will call this kind of relationship *physical* or *first-order resemblance*.[12] A representing vehicle and its represented object resemble each other at first order if they share physical properties, that is, if they are equal in some respects. For example, a colour chip—a small piece of card coated with coloured ink—is useful to interior designers precisely because it has the same colour as paint that might be used to decorate a room.

First-order resemblance, while relevant to certain kinds of public representation, is clearly unsuitable for connectionist representing vehicles, since it is incompatible with what we know about the brain's neural networks. Nothing is more obvious than the fact that our minds are capable of representing features of the world that are not replicable in patterns of neural activity. Moreover, even where the actual properties of neural networks are concerned, it is unlikely that these very often play a role in representing those self-same properties in the world.

There is, however, another kind of resemblance available. Consider colour chips again. Interior designers typically use *sets* of chips or colour *charts* to assist them in making design decisions. In other words, they employ a *system* of representations which depends on a mapping of paints onto chips according to their shared colour (their first-order resemblance). A useful side effect of having such a system is that when one wants to compare paints (e.g. 2-place comparisons such is "this one is bolder than that one", or 3-place comparisons such as "this one harmonises better

[11] This general approach to mental content has a venerable history in philosophy, but up until recently any kind of resemblance theory was thought to suffer from a number of fatal flaws (Cummins, 1989, Chap. 3). Over the last few years, however, a number of philosophers have started to take this approach seriously again, especially in the form of second-order resemblance relations (see especially Cummins, 1996).

[12] We are here adapting some terminology developed by Shepard and Chipman (1970).

with this one than with that one") one can do so by comparing the cards. This is because the system of chips embodies the *same pattern of colour-relations* as the paints. Whenever pairs or triples of paints satisfy particular colour relationships, their ink-coated proxies fall under mathematically identical relations.

Similar remarks apply to surface maps. What makes a map useful is the fact that it preserves various kinds of topographic and metrical information. The way this is accomplished is by so arranging the points on the map that when location A is *closer to* location B than location C, then their proxies (points A, B and C on the map) also stand in these metrical relations; and when location A is *between* locations B and C, then points A, B and C stand in the same (3-place) topographic relation; and so on. The utility of a map thus depends on the existence of a resemblance relation that assigns points on the map to locations in the world in such a way that the spatial relations among the locations is preserved in the spatial relations among the points.

We will speak here of *second-order resemblance*.[13] In second-order resemblance, the requirement that representing vehicles share physical properties with their represented objects can be relaxed in favour of one in which the *relations* among a system of representing vehicles mirror the *relations* among their objects. Of course, the second-order resemblance between colour charts and paints is a consequence of the first-order resemblance between individual chips and their referents. And in the case of surface maps, space is used to represent space. But one can typically imagine any number of ways of preserving the pattern of relations of a given system *without* employing first-order resemblance. For example, the height of a column of liquid in a mercury thermometer is used to represent the temperature of any object placed in close contact with it. Here, variations in height correspond to variations in temperature. And in a weather map the spacing of isobars is employed to represent pressure gradients, thus variations in isobar spacing mirror relations among pressure gradients (and wind velocities).

The significance of second-order resemblance for explaining the representational content of the brain's representing vehicles is this. While it is extremely unlikely that first-order resemblance is applicable to mental representation (given what we know about the brain) the same does not apply to second-order resemblance. Two systems can share a pattern of relations *without* sharing the physical properties upon which those relations depend. Essentially nothing about the physical properties of a system of representing vehicles is implied by the fact that it resembles a system of represented objects at second-order.

2.4. Second-order resemblance and connectionist vehicles

Second-order resemblance is arguably the right relation to explain the representational powers of connectionist representing vehicles. As an example consider

[13] See Palmer (1978), Shepard and Chipman (1970) and Shepard and Metzler (1971). Blachowicz (1997), Cummins (1996), Gardenfors (1996), Johnson-Laird (1983), O'Brien (1999) and Swoyer (1991) have all recently applied the concept of second-order resemblance to the problem of explaining representational content.

Cottrell's face-recognition network (see Churchland, 1995, pp. 38–55, for discussion). This network has a three layer feed-forward architecture: a 64×64 input array, fully connected to a hidden layer of 80 units, which in turn is fully connected to an output layer comprising 8 units. Each unit in the input layer can take on one of 256 distinct activation values, so it is ideal for encoding discretised grey-scale images of faces and other objects. After squashing through the hidden layer these input patterns trigger three units in the output layer that code for face/non-face status and gender of subject, and five which encode arbitrary 5-bit names for each of 11 different individuals. Cottrell got good performance out of the network after training it on a corpus of 64 images of 11 different faces, plus 13 images of non-face scenes. He found that the network was: (1) 100% accurate on the training set with respect to faceness, gender and identity (name); (2) 98% accurate in the identification of *novel* photos of people featured in the training set; and (3) when presented with entirely novel scenes and faces, 100% correct on whether or not it was confronting a human face, and around 80% correct on gender.

What is significant about the face-recognition network, for our purposes, is the way it codes faces at the hidden layer. Cluster analysis reveals that the network partitions its hidden unit activation space into face/non-face regions; within the face region into male/female regions; and then into smaller sub-regions corresponding to the cluster of patterns associated with each subject (see Fig. 1). Within the face region each point is an abstract (because compressed) representation of a face. Faces that are similar are represented by points that are close together in the space,

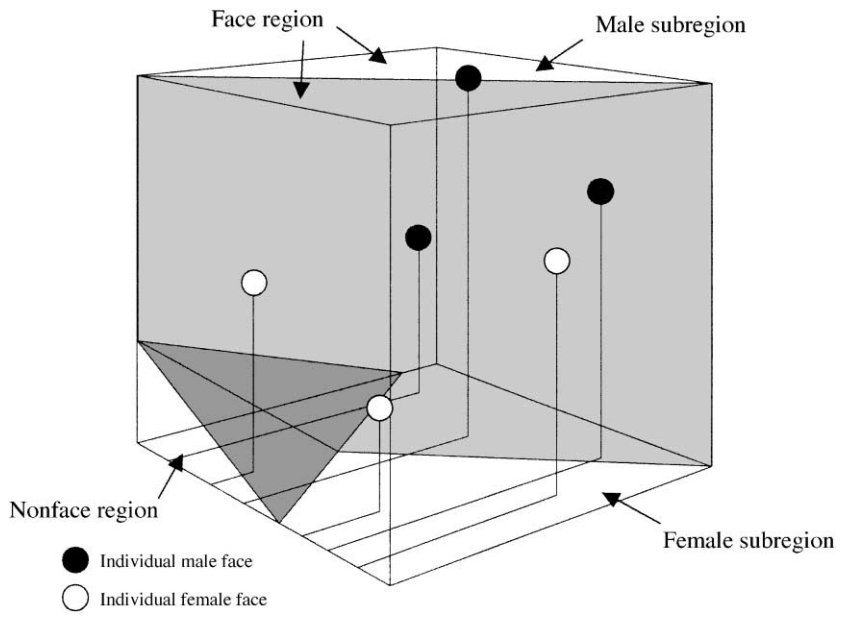

Fig. 1. The hierarchy of learned partitions across the hidden unit activation space of Cottrell's face recognition network (after Churchland, 1995, p. 49).

whereas dissimilar faces are coded by points that are correspondingly further apart. So the relations among faces which give rise to our judgments concerning similarity, gender, etc., are preserved in the distance relations in activation space.

Cottrell's face-recognition network thus appears to support a second-order resemblance between activation patterns and the domain of human faces. We can be more specific about the nature of this resemblance relation. Hidden unit activation space is a *mathematical* space used by theorists to portray the set of activation patterns a network is capable of producing over its hidden layer. Activation patterns themselves are physical objects (patterns of neural firing if realised in a brain), so distance relations in activation space actually codify *physical* relations among activation states. Consequently, the set of activation patterns generated across any implementation of the face-recognition network constitutes a system of representing vehicles whose physical relations capture relations among human faces. Let us refer to this variety of second-order resemblance—one based on the *physical* relations among a set of representing vehicles—as *structural* resemblance. One system structurally resembles another when the physical relations among the objects that comprise the first preserve aspects of the relational organisation of the objects that comprise the second.[14]

Structural resemblance underpins all the various examples of representation discussed in the last subsection. For example, the representing power of a mercury thermometer relies on a correspondence between one physical variable (the height of the column of mercury) and another (the temperature of bodies in contact with the thermometer). The significance of structural resemblance for connectionism is that it puts representational content right at the heart of cognition, by aligning it with the very properties that power the computational and behavioural capacities of connectionist networks. This is consistent with the connectionist focus on the intrinsic properties of representing vehicles, as opposed to their causal relations (see Section 2.2). Structural resemblance would thus appear to be the proper ground for representational content in connectionist computational systems.

2.5. Structural resemblance and abstract representational content

Structural resemblance is a form of second-order resemblance. In Section 2.3 we observed that the existence of a second-order resemblance relation between the brain's representing vehicles and some domain of represented objects implies nothing about the physical properties of those vehicles: a system of vehicles and a system of objects can resemble one another at second-order without sharing physical properties. What we didn't emphasise earlier is that a second-order resemblance relation

[14] Another variety of second-order resemblance is *functional* resemblance. A functional resemblance obtains when the pattern of *causal* relations among a set of representing vehicles mirrors the relations among a set of represented objects. This kind of resemblance is not appropriate for unpacking representation in connectionist systems. As we argued in Section 2.3, connectionism obliges us to explain representational content in terms of the *intrinsic* properties of the brain's representing vehicles. Functional resemblance doesn't do this; it focuses instead on the causal, and hence *extrinsic*, relations among a set of representing vehicles.

likewise implies nothing about the physical properties of the objects being represented. Indeed, second-order resemblance makes it possible for the brain's representing vehicles to resemble objects that don't possess *any physical properties at all*.

To make this clearer, let us be more precise about the nature of second-order resemblance. Suppose $S_V = (V, \mathcal{R}_V)$ is a system comprising a set V of objects, and a set $\mathcal{R}_V$ of relations defined on the members of V. We will say that there is a *second-order resemblance* between two systems $S_V = (V, \mathcal{R}_V)$ and $S_O = (O, \mathcal{R}_O)$ if, for at least *some* objects in V and *some* relations in $\mathcal{R}_V$, there is a one-to-one mapping from V to O and a one-to-one mapping from $\mathcal{R}_V$ to $\mathcal{R}_O$ such that when a relation in $\mathcal{R}_V$ holds of objects in V, the corresponding relation in $\mathcal{R}_O$ holds of the corresponding objects in O.[15] In other words, the two systems resemble each other with regard to their relational organisation. As already stressed, resemblance of this kind is independent of first-order resemblance, in the sense that two systems can resemble each other at second-order without sharing properties.

Second-order resemblance, so defined, is a very abstract relationship, not a substantial or physical one. The objects in V may be concrete or conceptual and the relations in $\mathcal{R}_V$ may be spatial, causal, structural, inferential, and so on. For example, V might be a set of features on a map, with various geometric and part-whole relations defined on them. Or V might be set of well-formed formulae in first-order logic falling under relations such as identity and consistency. A consequence of this is that a system of *mental* vehicles (which by assumption is a set of brain states) is not only capable of standing in a relationship of second-order resemblance to concrete or natural systems, but also to abstract systems such as logical formalisms and theories.

This fact about second-order resemblance, while infrequently remarked upon, has not gone completely unnoticed by theorists working in this area. Johnson-Laird, for example, has sought to extend his notion of a "mental model", which arguably is grounded in second-order resemblance, into the realm of abstract cognition (Johnson-Laird, 1983).[16] More recently, Blachowicz (1997) has suggested that the notion of analog representation, at least when this is understood under what he calls the "model interpretation" (again grounded in second-order resemblance, which he refers to as "relational identity"), can be taken out of its traditional context of perception and mental imagery, and used to illuminate the nature of our conceptual cognitive capacities.

[15] As defined, second-order resemblance is a relatively weak mapping. The literature on resemblance (e.g, Cummins, 1996, pp. 85–111) tends to focus on the far stronger notion of isomorphism. An isomorphism is a one-to-one, surjective (all to all), relation-preserving mapping. We suspect that where representation is concerned, the kind of mapping that is likely to be relevant will generally be weaker than isomorphism.

[16] Johnson-Laird writes at one point:

> Whenever I have talked about mental models, audiences have readily grasped that a layout of concrete objects can be represented by an internal spatial array, that a syllogism can be represented by a model of individuals and identities between them, and that a physical process can be represented by a three-dimensional dynamic model. Many people, however, have been puzzled about the representation of abstract discourse; they cannot understand how terms denoting abstract entities, properties, or relations can be similarly encoded, and therefore they argue that these terms can have only 'verbal' or propositional representations. (Johnson-Laird, 1983, p. 415)

If we are right that structural resemblance grounds the representational content of connectionist representing vehicles, what follows for radical connectionism? You will recall that this position, as against both classicism and ecumenical connectionism, denies that linguiform representational media play any part in human cognition. More particularly, radical connectionism denies that we think in our natural language. Given that symbolic representing vehicles are widely held to be the only road to abstract thought, this appears to create a problem for the radical connectionist. However, what the discussion above demonstrates is that connectionist systems are not precluded from representing the abstract merely because they eschew symbols. Any physical system capable of satisfying the constraints on second-order resemblance is thereby capable of representing objects that stand in logical, formal or conceptual relations. Connectionist devices achieve this not by acting as symbol processors, but by generating *analogs* of abstract objects: representing vehicles whose physical relations mirror the formal relations under which those objects fall.

Radical connectionism is thus able to handle representation of the abstract at least as well as its rivals. Given this, one might wonder whether natural language has any significant bearing on human cognition. In the next section we'll suggest that it does. Natural language has an important part to play in human thought, despite the fact that it doesn't constitute its representational medium.

3. Thinking with language

From the phenomenological perspective it isn't clear whether natural language is a representational medium of thought. On the one hand, we are constantly running words and sentences through our heads, even when performing quite trivial cognitive tasks. On the other, there is the familiar feeling that our thoughts are present in some form before we attempt to express them in natural language ("I know *what* I want to say, I just don't know *how* to say it"). Phenomenology simply doesn't settle the question. But phenomenology nonetheless provides us with a few clues. Most significantly, while it is not clear that natural language functions as a representational medium of thought, words and sentences certainly *accompany* many of our deliberations. More than this, words and sentences appear to play a *facilitating* role in the unfolding of our thoughts.

We'll develop this idea in what follows, arguing that natural language acts as a catalyst for cognition (especially more abstract cognition), which both organises and controls the computational activities of cognitive modules right across the brain. But first we need to consider the role that language plays in communication between brains.

3.1. "Natural languages are in the communication business, not the representation business"[17]

We argued above that structural resemblance grounds the representational content of connectionist representing vehicles. In recent work, Cummins goes further.

[17] Cummins (1996, p. 132).

He claims that resemblance (in the form of isomorphism—see footnote16) is the basis of *all* representation—that "representation is isomorphism" (Cummings, 1996, p. 109). For Cummins, this has the consequence that linguistic tokens, given that they are not isomorphic with the things they are interpreted to mean, are not representing vehicles. And this has the further consequence that natural language cannot be a representational medium of human cognition.

What is interesting about Cummins' position, for our purposes, is this. Cummins accepts that natural languages are a means by which humans communicate their thoughts to one another. But he rejects the traditional view that linguistic tokens do this by representing those thoughts. Instead, he takes natural languages to be conventional signalling schemes. Words and sentences, whether spoken or written, communicate my thoughts by triggering representing vehicles in you that encode similar thoughts (see especially, Cummings, 1996, pp.135–140). On this view of things, understanding what someone is saying is not a matter of comprehending the meaning of the communicative vehicles; it is a matter of recognising a speaker's intentions:

> Rather than a lexicon of expressions with their associated semantic properties (e.g. satisfaction conditions expressed in Mentalese), we have a lexicon of expressions paired with their governing conventions, these being, essentially, instructions for inferring the communicative intentions of their users. (Cummins, 1996, p. 140)

Words don't have meanings, according to Cummins; rather, they have the communicative function of triggering concepts (where the latter are to be understood as knowledge structures, not as abstract objects that act as the constituents of propositions). We often manage to communicate our thoughts, because we are party to lexical conventions that associate particular terms with particular concepts[18], and further (grammatical) conventions that permit sentences to be used as recipes for combining concepts into thoughts. But communication is successful only to the extent that the receiving brain embodies both the governing conventions, and the relevant knowledge structures: "Communication...works best among those who not only share a language but who share a lot of relevant knowledge as well"(Cummins, 1996, p. 141).

This communicative conception of natural language is echoed in a recent paper by Paul Churchland. He there develops a "neurosemantics" that has strong similarities with both Cummins' account and the story we developed in the previous section. But Churchland adds an interesting twist. He writes:

> Think of language, not so much as a system for representing the world, but as an acquired *skill*, both a motor skill and a perceptual skill. But do not think of

[18] This does *not* require that interlocutors internally represent, say, the fact that "ugly" is the term in their language associated with a whole lot of knowledge concerning ugliness. Rather, such conventions depend on communicative mechanisms, realised in individual brains, that simply trigger (somewhat idiosyncratic) concepts in response to linguistic input, and generate appropriate linguistic tokens in response to communicative intentions.

> it as the skill of producing and recognizing strings of words. Think of it instead as the acquired skill of perceiving...and manipulating...the brain activities of your conspecifics, and of being competent, in turn, to be the subject of reciprocal brain-manipulation. We don't usually think of a dinner-table conversation in these terms, but evidently that is what is going on. I am both following and steering your own cognitive activities, as you are both following and steering mine. (Churchland, in preparation, Sec. 8)

From this perspective, language not only has a role in communicating thoughts by triggering appropriate representing vehicles in target brains, it is also the means by which one brain can shape the cognitive activities occuring in another.

In the next section we will suggest that natural language plays these roles inside individual brains as well as between brains.

3.2. The internalisation of natural language

It was Vygotsky's great insight that after children acquire a natural language as a tool for communication, they "internalise" it, that is, they appropriate it as a cognitive tool (Vygotsky, 1962). But for Vygotsky, as for many later theorists (including those we are calling "ecumenical connectionists"), this process is one in which an external communicative scheme becomes an internalised *representational medium*: children learn to communicate with natural language, and then they learn to think in it. We fully agree with Vygotsky that natural language comes to play an important part in cognition. It's his understanding of this process that we question. We think the role that natural language plays internally is similar to the one it plays externally. That is, the internalisation of natural language is a process whereby a conventionally governed set of communicative signals is put to work inside a brain.

Consider the picture of communication we've been developing. Communication involves an exchange of signals between a source brain and a receiving brain. Such signals take the form of spoken, written or signed tokens with a particular physical shape (e.g. modulated sound waves, or ink marks on paper). They are produced when analog representing vehicles in the source brain interact with motor systems via complex mechanisms that realise the governing conventions of language. And they influence the receiving brain by impacting on its sensory surfaces, either directly (as with speech), or indirectly, by way of reflected photons (as with text, or expressions in a sign language).[19] Communication is successful when the emitted signals lead the receiving brain to token representing vehicles whose representational contents are sufficiently similar to those tokened in the source brain.

Cummins' insight is that linguistic signals need not (indeed *should* not) be conceived as content-bearers in order to explain their role in this process. Churchland's further insight is that such signals fundamentally operate as a means by which we

[19] Whether the impact of a signal is direct or indirect, some processing in the receiving brain is always required to recover low level lexical features (such as phonemes or graphemes).

manipulate the contents and trajectory of thought in other people. What Vygotsky adds to the mix is the idea that natural language gets internalised during development—it becomes a system of signals apt not only for manipulating the brains of others, but also for recurrent *self*-manipulation.[20] Such internalisation involves the establishment and maintenance of causal/communicative links across a single brain.

The process of communication, since it begins and ends with representing vehicles, can be internalised insofar as: (1) some internal state is able to "stand in" for the external signal, and (2) the brain's internal cognitive economy can be so arranged that this internal state stands in a similar causal relation to thoughts as does the external signal. The first condition is relatively easy for a brain to satisfy, since brains are in the business of constructing internal models of external objects and states of affairs. That is, the obvious internal analog of an external signal is a representing vehicle that takes the signal as its represented object.[21] And if the signal is internalised in the form of a representing vehicle, it should be possible to arrange matters so that it stands in the requisite causal relations with the vehicles that code the communicated thought, thus satisfying the second condition.

At this juncture one might wonder about the point of internalising this process. What good is a thought that generates an internalised signal that then generates another thought, all in the one brain? The good is this: once a brain has internalised a set of conventionally governed signals, these signals can be employed by one part of the brain to steer the cognitive activities occurring in other parts of the brain.[22] Natural language thereby becomes a powerful cognitive tool; one that can establish coherent, multi-modal representational states involving many brain sites, by facilitating communication among those sites; and one that can regulate the sequencing of thought, via the constant interplay between networks that encode linguistic signals and those that encode thoughts. There is emerging evidence that language, implemented primarily in temporal cortex, plays just these roles (for discussion see Damasio, 1989, 1994). Recurrence, in the form of causal processes that loop from language centres out to the analog representing vehicles they trigger, and back again, plays a crucial role in all of this. Such causal loops catch up language and thought in a tight web of mutual influence that extends our cognitive capacities well beyond those of infra-verbal organisms.

We have defended radical connectionism. Radical connectionism claims, as against both classicism and ecumenical connectionism, that cognition *never* involves an internal symbolic medium, not even when natural language plays a part in our thought processes. On the face of it, this renders the human capacity for abstract thought quite mysterious. However, we've argued that connectionism, because it

[20] Cognitive self-manipulation need be no more involved than talking (out loud) to oneself. During internalisation, overt egocentric speech falls away, to be replaced by inner speech.

[21] One should not assume that such a representing vehicle carries a representational content equivalent to the conventional meaning of the signal. To do so would be to hold that the system of internalised signals constitutes a symbolic representational medium. The content of the representing vehicle is the physical object that constitutes the external signal—e.g. an uttered or written word or sentence.

[22] This idea is somewhat similar to (but not identical with) the speculations that Dennett makes about the role of natural language in organising our thinking (see, e.g Dennett, 1991, pp. 193–199)

adopts an analog conception of neural computation, is committed to a structural resemblance theory of representational content. Representation of the abstract is no more problematic for a system of analog vehicles that structurally resemble their target domain, than for a symbol system. Natural language is therefore not required as a representational medium for abstract thought. Indeed, since natural language is arguably not a representational medium *at all*, but a conventionally governed scheme of communication, the role of internalised (i.e. self-directed) language is best conceived in terms of the coordination and control of cognitive activities within the brain.

References

Bechtel, W., Abrahamsen, A.A., 1991. Connectionism and the Mind: An Introduction to Parallel Processing in Networks. B. Blackwell.

Blachowicz, J., 1997. Analog representation beyond mental imagery. Journal of Philosophy 94 (2), 55–84.

Churchland, P.M., 1995. The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain. MIT Press, Cambridge, MA.

Churchland, P. M., in preparation. Neurosemantics: On the mapping of minds and the portrayal of worlds.

Churchland, P.S., Sejnowski, T.J., 1992. The Computational Brain. MIT Press, Cambridge, MA.

Clapin, H., O'Brien, G., 1998. A conversation about superposition and distributed representation. Noetica: Open Forum 3 (10).

Clark, A., 1989. Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing. MIT Press, Cambridge, MA.

Clark, A., 1993. Associative Engines: Connectionism, Concepts, and Representational Change. MIT Press, Cambridge, MA.

Clark, A., 1997. Being There: Putting Brain, Body and World Together Again. MIT Press, Cambridge, MA.

Cummins, R., 1989. Meaning and Mental Representation. MIT Press, Cambridge, MA.

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

Damasio, A.R., 1989. Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. Cognition 33 (1–2), 25–62.

Damasio, A.R., 1994. Descartes' Error: Emotion, Reason, and the Human Brain. G.P. Putnam.

Dennett, D., 1982. Styles of mental representation. Proceedings of the Aristotelian Society, New Series 83, 213–226.

Dennett, D.C., 1991. Consciousness Explained. Little, Brown.

Devitt, M., Sterelny, K., 1987. Language and Reality. Blackwell, Oxford.

Dretske, F., 1981. Knowledge and the Flow of Information. Blackwell, Oxford.

Field, H., 1978. Mental representation. Erkenntnis 13, 9–61.

Fodor, J., 1975. The Language of Thought. Harvester Press, Cambridge, MA.

Fodor, J., 1987. Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press, Cambridge, MA.

Fodor, J., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.

Gardenfors, P., 1996. Mental representation, conceptual spaces and metaphors. Synthese 106 (1), 21–47.

Harman, G., 1973. Thought. Princeton University Press, Princeton, N.J.

Johnson-Laird, P.N., 1983. Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, Cambridge, MA.

Lloyd, D., 1989. Simple Minds. MIT Press, Cambridge, MA.

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

O'Brien, G., 1998. The role of implementation in connectionist explanation. Psycoloquy 9 (6).

O'Brien, G., 1999. Connectionism, analogicity and mental content. Acta Analytica 22, 111–131.

O'Brien, G., Opie, J., 1999. A connectionist theory of phenomenal experience. Behavioral and Brain Sciences 22 (1), 127–148.

Opie, J., 1998. Connectionist modelling strategies. Psycoloquy 9 (30).

Palmer, S., 1978. Fundamental aspects of cognitive representation. In: Rosch, E., Lloyd, B. (Eds.), Cognition and Categorization. Lawrence Erlbaum, Hillsdale, N.J.: New York.

Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.

Rumelhart, D.E., Smolensky, P., McClelland, J.L., Hinton, G.E., 1986. Schemata and sequential thought processes in PDP models. In: McClelland, J.L., Rumelhart, D.E. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2. MIT Press, Cambridge, MA.

Sejnowski, T., Rosenberg, C., 1987. Parallel networks that learn to pronounce English text. Complex Systems 1, 145–168.

Shepard, R., Chipman, S., 1970. Second-order isomorphism of internal representations: shapes of states. Cognitive Psychology 1, 1–17.

Shepard, R.N., Metzler, J., 1971. Mental rotation of three-dimensional objects. Science 171 (972), 701–703.

Smolensky, P., 1987. The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn. Southern Journal of Philosophy 26, 137–161.

Smolensky, P., 1988. On the proper treatment of connectionism. Behavioural and Brain Sciences 11, 1–23.

Swoyer, C., 1991. Structural representation and surrogative reasoning. Synthese 449–508.

Tienson, J., 1987. Introduction to connectionism. Southern Journal of Philosophy 26, 1–16.

van Gelder, T., 1990. Compositionality: a connectionist variation on a classical theme. Cognitive Science 14 (3), 355–384.

Von Eckardt, B., 1993. What is Cognitive Science? MIT Press, Cambridge, MA.

Vygotsky, L., 1986. Thought and Language. MIT Press, Cambridge, MA.