# A CONNECTIONIST DEFENCE

# OF

# THE INSCRUTABILITY THESIS

# AND

# THE ELIMINATION OF THE MENTAL

FRANCISCO JOSÉ CALVO GARZÓN

DOCTORAL DISSERTATION

UNIVERSITY OF GLASGOW

DEPARTMENT OF PHILOSOPHY

DECEMBER, 1999

This thesis is submitted in fulfilment of the requirements for the degree of Doctor in Philosophy (Ph.D.) at the University of Glasgow, Faculty of Arts, Department of Philosophy.



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

# Abstract

This work consists of two parts. In Part I (chapters 1—5), I shall produce a Connectionist Defence of Quine's Thesis of the Inscrutability of Reference, according to which there is no objective fact of the matter as to what the *ontological commitments* of the speakers of a language are. I shall start by reviewing Quine's project in his original behaviouristic setting. Chapters 1, and 2 will be devoted to addressing several criticisms that Gareth Evans, and Crispin Wright, have put forward on behalf of the friend of *semantic realism.* Evans (1981) and, more recently, Wright (1997) have argued on different grounds that, under certain conditions, *structural* simplicity may become alethic—i.e., truth-conducive—for semantic theories. Being structurally more complex than the standard semantic theory, Quine's perverse semantic route (see chapter 1) is an easy prey for Evans' and Wright's considerations. I shall argue that both Evans' and Wright's criticisms are unmotivated, and do not jeopardize Quine's overall enterprise. I shall then propose a perverse theory of reference (chapter 3) which differs substantially from the ones advanced in the previous literature on the issue. The motivation for pursuing a different perverse semantic proposal resides in the fact that the route I shall be offering is as simple, structurally speaking, as our sanctioned theory of reference is meant to be. Thanks to this feature, my strategy is not subject to certain criticisms which may put perverse proposals à la Quine in jeopardy, thereby becoming an overall better candidate for the Quinean to fulfill her goal. In

chapter 4, I shall introduce and develop a criterion recently produced by Wright (1997) in terms of 'psychological simplicity' which threatens the perverse semantic proposal I offered in chapter 3. I shall argue that a Language-of-Thought—LOT—model of human cognition could motivate Wright's criterion. I shall then introduce the reader to some basic aspects of connectionist theory, and elaborate on a particularly promising neurocomputational approach to language processing put forward by Jeff Elman (1992; 1998). I shall argue that if instead of endorsing a LOT hypothesis, we model human cognition by a *recurrent* neural network à la Elman, then Wright's criterion is unmotivated. In particular, I shall argue that considerations regarding 'psychological simplicity' are *neutral*, favouring neither a standard theory of reference, nor a perverse one. In the remainder of Part I, I shall focus upon certain problems for the defender of the Inscrutability Thesis highlighted by the friend of connectionist theory. In chapter 5 I shall introduce a mathematical technique for measuring conceptual similarity across networks that Aarre Laakso and Gary Cottrell (1998; 2000) have recently developed. I shall show how Paul Churchland makes use of Laakso and Cottrell's results to argue that connectionism can furnish us with all we need to construct a robust theory of semantics, and a robust theory of translation—robustness that may potentially be exploited by a connectionist foe of Quine to argue against the Inscrutability Thesis. The bulk of the chapter will be devoted to showing that the notion of conceptual similarity available to the connectionist leaves room for a "connectionist Quinean" to kick in with a one-to-*many* translational mapping across networks.

In Part II (chapters 6, and 7), I shall produce a Connectionist Defence of the Thesis of Eliminative Materialism, according to which propositional attitudes don't exist (see chapter 7). I shall start by rejoining to two arguments that Stephen Stich has recently put forward against the thesis of eliminative materialism. In a nutshell, Stich (1990; 1991) argues that (i) the thesis of eliminative materialism, is neither true nor false, and that (ii) even if it were true, that would be philosophically uninteresting. To support (i) and (ii) Stich relies on two premises: (a) that the job of a theory of reference is to make

explicit the tacit theory of reference which underlies our intuitions about the notion of reference itself; and (b) that our intuitive notion of reference is a highly idiosyncratic one. In chapter 6 I shall address Stich's anti-eliminativist claims (i) and (ii). I shall argue that even if we agreed with premises (a) and (b), that would lend no support whatsoever for (i) and (ii). Finally, in chapter 7, I shall introduce a connectionist-inspired conditional argument for the elimination of the posits of folk psychology put forward by William Ramsey, Stephen Stich, and Joseph Garon. I shall consider an objection to the eliminativist argument raised by Andy Clark. I shall then review a counter that Stephen Stich and Ted Warfield produce on behalf of the eliminativist. The discussion in chapter 5 on 'state space semantics and conceptual similarity' will be used to show that Clark's argument is not threatened by Stich and Warfield's considerations. Then, in the remainder of Part II, I shall offer a different line of argument to counter to Clark. A line that focuses on the notion of causal efficacy. I hope to show that the thesis of eliminativist materialism is correct. Conclusions, and directions for future research will follow.

For Ana, *El Pollico*

*She had always wanted words, she loved them,*
*grew up on them. Words gave her clarity,*
*brought reason, shape. Whereas I thought*
*words bent emotions like sticks in water.*

—Michael Ondaatje, *The English Patient*

# Contents

List of Figures viii

Acknowledgements xii

## Part I

*The Inscrutability of Reference*

**Chapter 1**

*The Inscrutability of Reference*

**1.1** *Introduction* *1*

**1.2** *Levels of Explanation: Behaviourism and/or Neurophysiology* *5*

**1.3** *First Steps in Radical Translation: Stimulus Meaning and Observation Sentences* *11*

**1.4** *The Juggling Strategy* *16*

**1.5** *Locating Schemes of Predication* *23*

**1.6** *Evans' First Counter-Example, and Hookway's 'Divide-and-Rule' Strategy* *30*

**1.7** *Evans' Second Counter-Example: Widening the Scope of Hookway's Disjunctive Route* *35*

**1.8** *Conclusion* *41*

**Chapter 2**

*Is Simplicity Alethic for Semantic Theories?*

**2.1** *Introduction* *43*

**2.2** *Evans' Mirror Constraint* *46*

**2.3** *Full-blooded Semantics and Disjunctive Semantic Theories* *50*

**2.4** *Wright's "Metodological Simplicity" Criterion* *55*

**2.5** *Beefing up Semantic Discourse: A Reductio contra Wright* *58*

**2.6** *The Metaphysical Status of Semantics* *66*

**2.7** *Conclusion* *69*

**Chapter 3**

*Semantic Perversity*

**3.1** *Introduction* *71*

**3.2** *A Perverse Way of Dividing Reference over Parts of Things* *73*

**3.3** *The '99%-urp' Scheme and Evans' Second Counter-Example* *79*

**3.4** *Back to the Apparatus of Individuation* *82*

**3.5** *Game-Theoretical Semantics* *86*

**3.6** *Conclusion* *96*

**Chapter 4**

*A Connectionist Defence of the Inscrutability Thesis*


| | | |
|---|---|---|
| **4.1** | *Introduction* | *98* |
| **4.2** | *Wright's 'Psychological-Simplicity' Argument* | *100* |
| **4.3** | *Psychological Simplicity and The Syntactic Image* | *105* |
| **4.4** | *Basic Aspects of Connectionism: Components and Dynamics* | *109* |
| **4.5** | *Learning and Conceptual Organization in Neural Networks: State Space Semantics* | *114* |
| **4.6** | *Simple Recurrent Networks and Conceptual Inclusion* | *126* |
| **4.7** | *Statistical Analyses, Symbolic Approximation, and Causal Efficacy* | *138* |
| **4.8** | *Systematicity, Compositionality, and the Generality Constraint* | *159* |
| **4.9** | *Conclusion* | *172* |


**Chapter 5**

*State Space Semantics and Conceptual Similarity*


| | | |
|---|---|---|
| **5.1** | *Introduction* | *179* |
| **5.2** | *State Space Semantics: The Problem* | *182* |
| **5.3** | *A Connectionist Measure of Conceptual Similarity* | *186* |
| **5.4** | *Similarity of Prototypical Trajectories: A Solution?* | *190* |
| **5.5** | *A Connectionist Approach to Radical Translation: First Reply to Churchland* | *195* |
| **5.6** | *The Collateral Information Challenge: Second Reply to Churchland* | *208* |
| **5.7** | *Conclusion* | *218* |

# Part II

## *The Elimination of the Mental*

**Chapter 6**

*Can We Turn a Blind Eye to Eliminativism?*

**6.1** *Introduction* *221*

**6.2** *Eliminativism and Folk Semantics* *223*

**6.3** *Rejoinder to Stich's First Argument* *225*

**6.4** *Eliminativism and the Idiosyncracy of Reference* *232*

**6.5** *Rejoinder to Stich's Second Argument* *235*

**6.6** *Conclusion* *243*

**Chapter 7**

*Connectionism and the Twilight of Propositional Content*

**7.1** *Introduction* *245*

**7.2** *Propositional Modularity and Fully-Superposed Neural Networks* *247*

**7.3** *Higher Levels of Description, NETtalkers, and NETtalk-structures* *259*

**7.4** *Cluster Analysis, and Causal Efficacy* *278*

**7.5** *Conclusion* *284*

**Bibliography** *296*

# Figures

Fig. 4.1 The sigmoid activation function often used for units in neural networks. Outputs (along the ordinate) are shown for a range of possible inputs (abscissa). Units with this sort of activation function exhibit an all or nothing response given very positive or very negative inputs; but they are very sensitive to small differences within a narrow range around 0. With an absence of input, the nodes output 0.5, which is in the middle of their response range. (from Elman *et al.*, 1996, p. 53)

Fig. 4.2 A simple feedforward network. Units in layer 1 are input units. Units in layer 3 are output units. Activations strictly feed forward. Each unit in layer *n* has output connections to each unit in layer *n+1*.

Fig. 4.3 The activation-vector space of the middle layer of the acoustic network for sonar analysis. Note the partition into two exclusive categories: mine echoes and rock echoes. Note also the two prototypical hot spots where typical and uncompromised examples of each category are routinely coded. (from Churchland, 1995, p. 83)

Fig. 4.4 Elman's recurrent network used to discriminate grammatically correct sentences. (Elman, 1992, p. 153)

Fig. 4.5 Trajectories through state space for'[boy who boys chase chases

boy' and 'boys who boys chase chase boy']. After the indicated word has been input, each point marks the position along the second principal component of hidden unit space. Magnitude of the second principal component is measured along the ordinate; time (i.e., order of words in sentence) is measured along the abscissa. [...The] sentence-final word is marked with a ]s. (Adapted from Elman, 1992, pp. 162-3)

Fig. 4.6 Hierarchical clustering of hidden unit activation patterns from the sentence-prediction task (Elman, 1990). The network learns distributed representations for each word which reflects its similarity to other words. Words and groups of words which are similar are close in activation space, and close in position in the tree. (Elman *et al.*, 1996, p. 96)

Fig. 4.7 The finite state grammar deployed by Reber (1967; 1976). Numerals indicate states, and letters indicate transition arcs. Sentences are generated by traversing a path from initial state #0 to final state #5. After each transition a letter is produced, obtaining thus sequentially a string. (From Servan-Schreiber *et al.*, 1988, p. 6)

Fig. 4.8 General architecture of the network. (from Servan-Schreiber *et al.*, p. 7)

Fig. 4.9 Hierarchical Cluster Analysis of the H.U. activation patterns after 200,000 presentations from strings generated at random according to the Reber grammar (Three hidden units). (From Servan-Schreiber *et al.*, p. 10)

Fig.4.10 Hierarchical Cluster Analysis of the H.U. activation patterns after 200,000 presentations from strings generated at random according to the Reber grammar (Fifteen hidden units). (From Servan-Schreiber *et al.*, p. 14)

Fig. 5.1 Symmetric matrices obtained by taking Euclidean distances between all the representations in each network. (From Laakso and Cottrell, 2000)

Fig. 5.2 Trajectories through state space for '[boy who boys chase chases boy' and 'boys who boys chase chase boy']. After the indicated word has been input, each point marks the position along the second principal component of hidden unit space. Magnitude of the second principal component is measured along the ordinate; time (i.e., order of words in sentence) is measured along the abscissa. [...The] sentence-final word is marked with a ]s. (Adapted from Elman, 1992, pp. 162-3)

Fig. 7.1 A semantic network representation of memory in the style of Collins and Quillian (1972). (From RS&G, 1995, p. 318)

Fig. 7.2 Semantic netwotk with one proposition removed. (*Ibid.*, p. 319).

Fig. 7.3 Weights and biases in network with 16 propositions. (*Ibid.*, p. 326)

Fig. 7.4 Weights and biases in network with 17 propositions. (*Ibid.* p. 328)

Fig. 7.5 Hierarchy of partitions on hidden-unit vector space of NETtalk. (From Churchland, 1989, p. 176, after Rosenberg and Sejnowski, 1987)

Fig. 7.6 The locations of four prototype points within the hidden-layer activation spaces of two (imaginary) neural networks for recognizing the faces of four different extended families. The four points represent a prototypical Hatfield face, a prototypical McCoy face, a prototypical Wilson face, and a prototypical Anderson face. (Churchland, 1998, p. 9)

Table 7.1 Propositions Network A and Network B. (*Ibid.*, p. 324)

# Acknowledgements

In writing this dissertation, I have had the benefit of many helpful comments and suggestions from a number of people. For this, I would like to thank Bill Casebeer, Paul Churchland, Gary Cottrell, Jeff Elman, Patricia Kitcher, Aarre Laakso, and Philippe Schyns. Most of all, I am grateful to my thesis supervisor, Jim Edwards, for his help, encouragement, and unparalleled support provided during the four years that I was his student.

Some of the material of this work has been presented at the Graduate Seminars at the Department of Philosophy (Univeristy of Glasgow), and at conferences elsewhere. In particular, I would like to thank Bob Hale, Philip Percival, Pat Shaw, and Crispin Wright for very helpful comments and suggestions on preliminary versions of chapters 2, 3, and 4, delivered at the *Joint Philosophy Conferences* of the Department of Logic and Metaphysics (University of St. Andrews), and the Department of Philosophy (University of Glasgow) at the Isle of Raasay (Scotland) in the falls of 1996, and 1997. I would also like to thank audiences at the Departments of Cognitive Science, and Philosophy (University of California, San Diego) for discussion of versions of chapters 4, and 5, delivered at the *EPL Meetings* (University of California, San Diego).

Parts of chapters 1—5 are based on the following articles of mine that are forthcoming in scholarly journals:

"A Connectionist Defence of the Inscrutability Thesis", *Mind and Language*, (2000).

"State Space Semantics and Conceptual Similarity: Reply to Churchland", *Philosophical Psychology* 13, pp. 77-95 (2000).

"Is Simplicity Alethic for Semantic Theories?", *Analysis*, (under review).

"Semantic Perversity", *Teorema*, (under review).

The work reported here was supported by a variety of funding agencies. My research was supported by grants from:

*Caja de Ahorros del Mediterráneo—British Council,* *1995-97*
(University of Glasgow)
*Institute of Cell and Molecular Biology (University of Edinburgh), and Darwin Trust of Edinburgh,* *1997-98*
(University of Glasgow)
*Fulbright Fellowship (funded by CajaMurcia),* *1998-99*
(University of California, San Diego)
*and,*
*the Sir Richard Stapley Educational Trust.*

I would also like to thank a good friend of mine, Onofre T. Martínez Salinas, for his availability (late at night) to help me out with the formatting and printing of the final version. I'm also thankful to the secretaries of the Department of Philosophy, Anne and Frances, for their kindness and technical support. To my parents *Obe* and *Pula* to whom I owe everything in life, and to my sisters *Maena* and *Pingo* (*Peter Pan of Carrasca*) for their love. And last, but not least, my deepest debt goes to my wife Ana and my baby girl-to-be, Hortensia, without whom not.

# Part I

## *The Inscrutability of Reference*

---

*El mundo era tan reciente, que muchas cosas carecían de nombre, y para mencionarlas había que señalarlas con el dedo.*

—Gabriel García Márquez, *Cien Años de Soledad*

# 1

# *THE INSCRUTABILITY OF REFERENCE*

## 1.1 *Introduction*

In a nutshell, Quine's Thesis of the Inscrutability of Reference claims that there is no objective fact of the matter as to what the *ontological commitments* of the speakers of a language are (see below). To become acquainted with this polemical thesis, Quine (1960) invites the reader to imagine two linguists whose task is to produce rival translation manuals to account for the expressions of an unknown language. The linguists, as we shall see shortly (1.3, and 1.4 below), can produce rival translation manuals which are mutually incompatible, and yet fit all possible evidence.[1] The Inscrutability Thesis is the doctrine that there is no fact of the matter as to what the extensions of the terms of a language are. Claims about the ontological commitments that the speakers of a language incur are relative to which

[1] The question of just what source of evidence should be taken into account must wait until section 1.2.

translation manual we favour. Were this radical thesis to earn its keep, objectivism as applied to our ordinary notion of reference, and related semantic notions—truth, meaning, etc.—would be in serious jeopardy.

Quine's most famous illustration of the Inscrutability Thesis comes from the 'gavagai' parable (see 1.3 below).[2] Unfortunately, the 'gavagai' parable has contributed to the proliferation of disparate interpretations as to what thesis it is meant to exemplify. Although I shall not seek to elucidate exegetical issues, let me introduce a small caveat before kicking off that will help us clarify a potential source of misunderstandings, and what the scope of the present work is.

The reader of *Word and Object* will recall the 'gavagai' example as an illustration of the Thesis of the Indeterminacy of Translation. Although intimately related, the Indeterminacy of Translation, and the Inscrutability of Reference make

---

[2] Throughout his writings, Quine has made use of a number of examples to illustrate the inscrutability of reference. In 'Ontological Relativity', for instance, Quine mentions in support of his thesis certain Japanese syntactic constructions, called *classifiers*. This constructions can be translated into English either as mass terms, or as individuative terms (see Quine, 1969a, pp. 35-9, for a detailed explanation). Unfortunately, cases like the Japanese classifiers, have lead to some friends of semantic realism to fall into the temptation of mistaking the Inscrutability Thesis for a platitude. In the case of Japanese classifiers, referential inscrutability is achieved by exploiting syntactic or semantic resources that one language has, and the other language lacks. However, this can only bring referential inscrutability in a trivial way. Quine's thesis is meant to obtain even when the languages under consideration do share the same syntactic and semantic apparatus. Indeed, referential inscrutability, according to Quine, is supposed to hold at Home, among fellow speakers (the reader may care to consult Kirk, 1986, chapter 5, for an insightful discussion of this issue. See also Field, 1975, p. 396). In my defence of the Inscrutability Thesis in Part I of the present work, I shall focus exclusively upon the 'gavagai' parable; example which seems not to lead to the trivializations that Japanese classifiers, and other examples that Quine deploys, have lead some commentators to.

different contentions, and by contrast to what many believe, only the latter receives support from Quine's parable of Radical Translation. Put bluntly, the Indeterminacy of Translation claims that two linguists could produce rival manuals of translation to account for the sense of the sentences of a foreign language. The Indeterminacy Thesis deals with whole sentences, and targets the alleged determinacy of *sense* pursued by the semantic realist by questioning the factuality of the semantic relations of synonymy and translation. By contrast, the Inscrutability Thesis deals with *terms*, and targets the alleged scrutability of *reference*—cf. Quine, 1969a, pp. 34-ff. Both theses target objectivism in semantics, although at different levels. The Indeterminacy Thesis highlights scepticism about the *museum myth*. That is, the existence of meanings as *mental items* to which words are assigned. The Inscrutability Thesis, on the other hand, highlights a deeply interwined, although different, myth about reference. Namely, the idea that words can be attached to things via *mental acts*, such that for example the word 'rabbit' gets connected to *rabbits*. As I said, and contrary to common wisdom, the 'gavagai' example only exemplifies the Inscrutability Thesis—cf. Quine, 1970d.[3]

Moreover, there has been a lot of controversial debate about whether the Inscrutability Thesis—were it to be sound—can bring support (directly, or indirectly) to the Indeterminacy of Translation. In 'On the Reasons for Indeterminacy of Translation' Quine is pretty explicit about it, claiming that the Inscrutability Thesis does not entail indeterminacy at the sentential level. The

---

[3] To many readers it is not obvious how we can have inscrutability of reference without indeterminacy of sentence translation. See fn. 20 below for some clarifying remarks that Quine makes in that respect; remarks that must await until the 'gavagai' parable is reviewed in sections 1.3, and 1.4.

Indeterminacy of Translation receives support from another important thesis of Quine: The Thesis of Undermination of Scientific Theories by All Possible Observation.[4] Recent commentators such as Kirk (1986), however, argue that the Inscrutability Thesis does entail indeterminacy of sentence translation.[5] Granting that entailment, nonetheless, should not cause any concern to the foe of Quine. Kirk argues indirectly against the Inscrutability Thesis, claiming that (i) inscrutability of reference entails indeterminacy of translation, and that (ii) there is no indeterminacy of translation. But, obviously, someone's *tollens* is somebodyelse's *ponens*. In the present work, I shall focus exclusively upon the Inscrutability of Reference, and try to show that it is correct. In an opposite direction to Kirk's line of reasoning, it may then be argued that the Indeterminacy of Translation holds too—were my defence of the Inscrutability Thesis to succeed. Nevertheless, the purpose of Part I of my dissertation is more modest, and I shall not address Kirk's alleged connection between reference of terms, and translation at the sentential level.

Before getting started let me briefly outline the programme of this chapter. In section 1.2 I shall introduce two caveats regarding the evidential basis, and the reading (ontological *versus* epistemological) that we should follow when reading Quine's Inscrutability Thesis. An appraisal of these two caveats will help clarify the main contention, and scope of my defence of Quine's thesis. In sections 1.3, and 1.4 I shall review Quine's project of Radical Translation in his original behaviouristic

---

[4] See Quine (1970d; 1975a); and Wright (1997) for an insightful appraisal of that thesis, and its putative bearing upon the Indeterminacy of Translation.

[5] See also Levin, 1979, p. 25; and Davidson, 1984, p. 227, for a positive link between both theses.

setting.[6] In section 1.5 I shall pave the way for an appraisal of a criticism put forward by Evans in his seminal paper 'Identity and Predication'. Sections 1.6, and 1.7 will be devoted to analyzing two counter-examples that Evans offers against Quine's Inscrutability Thesis. I shall exploit a strategy for dealing with Evans' first counter that, on behalf of Quine, Hookway (1988) has developed. However, there is a further counter produced by Evans which Hookway doesn't address. In section 1.7 we'll see how a manual, along the lines of the one Hookway advances, manages to overcome these further difficulties raised by Evans. Conclusions will follow in section 1.8.

## 1.2 *Levels of Explanation: Behaviourism and/or Neurophysiology*

The lack of factuality regarding what the terms of a given language refer to is a claim that carries a lot of weight, and requires at least two important clarifications before we can review Quine's parable of Radical Translation. On the one hand, it seems that such a strong contention must depend at least partly upon what sort of evidence we take to be relevant to the fact of the matter. In this respect, Quine's approach to semantics is crucially shaped by his naturalism, and the behaviourism dominant in philosophy, and the experimental sciences in the 50s. All aspects of human activity must be studied under the light of Natural Science. And this includes, of course, linguistic activity. In the opening passages of 'Ontological Relativity', Quine quotes Dewey, to whom he owes academically his inclination towards naturalism: 'Meaning...is not a psychic existence; it is primarily a property

[6] The reader familiar with Quine's parable (*Word and Object*, chapter 2) may wish to skip these two sections, and jump ahead to section 1.5.

of behavior' (Quine, 1969a, p. 27). Quine will adopt this commandment as the basis for his analysis of the semantic notion of reference. As a result, Quine proclaims, *all* the objective evidence we can make use of in our search reduces to *behavioural* evidence—in particular, evidence about people's behavioural (linguistic and non-linguistic) dispositions (see 1.3 below). Plausibly, however, by admitting a richer evidential basis beyond people's behavioural dispositions, the room for scepticism towards semantic factuality might be significantly reduced.[7]

Quine (1975c) distinguishes three possible levels of explanation for dealing with semantic issues: the mentalistic, the behaviouristic and the neurophysiological. The mentalistic level comprehends among others, facts about beliefs, desires, and the rest of the propositional attitudes. Many philosophers have argued that data involving the intentional (mentalistic) apparatus of the speakers of a language reports matters of fact. Quine agrees that if the mentalistic level of explanation counted as among the matters of fact that constitute the genuine evidential basis for the linguist, translation would be determined, and reference would be scrutable. The reader may recall for example Lewis' (1974) principles of charity, rationality, etc., that act as a filter for putative ascriptions to the natives' beliefs, desires and others intentional attitudes they might have towards their environment.[8] However, Quine (1960, p. 221) denies the alleged factuality of the propositional attitudes. Beliefs,

---

[7] One of the first and more important reactions to Quine's restricted behaviouristic picture is due to Chomsky (1969; 1975). Quine's reply appears in Davidson, and Hintikka (1969). The failures of behaviourism are well-known and I shall not address them here. The reader may consult for example Kim (1996) for an overview of the main reasons for the collapse of behaviourism.

[8] Although see Davidson (1984) for reasons on why such ascriptions do not make translation determinate or reference scrutable.

desires, and other mentalistic idioms are fine, insofar as we grasp them in a *practical* way.[9] Once we assume for argument's sake that the mentalistic level is to be ruled out as part of a scientific inquiry into semantics, the question is: Can naturalism admit other relevant facts beyond the behavioural dispositions of speakers? At this point, my defence of the Inscrutability Thesis will differ from Quine's original setting. Quine's approach to the project of Radical Translation is behaviouristic in spirit:

> (our) talk of external things, our very notion of things, is just a conceptual apparatus that helps us to foresee and control the triggering of our sensory receptors in the light of previous triggering of our sensory receptors. The triggering, first and last, is all we have to go on. (Quine, 1981a, p. 193).

Quine's reasons to work at the medium (behaviouristic) level spring from his overall effort to naturalize epistemology (see Quine, 1969a, chapter 3) and, at the same time, from considerations concerning the process by which ordinary speakers acquire their mother language (see Quine, 1970b). Apparently, however, there is no reason why a naturalistic approach to semantics could not observe what goes on inside people's heads in addition to observing linguistic, as well as non-linguistic, behavioural dispositions. If we are to endorse a full-blooded naturalistic reading of semantics, I contend, the neurophysiological level of explanation must be taken into account. Mental phenomena, I claim, are not actually reducible to/replacable by the behavioural level, but rather by the neurophysiological level underlying behavioural

---

[9] I shall not review Quine's (behaviouristic) reasons for rejecting the propositional attitudes. In Part II of my dissertation (chapters 6, and 7) I shall argue, on grounds different from Quine's, against the propositional attitudes, producing a Connectionist Defence of the Elimination of the Mental.

dispositions.[10] In fairness to Quine, this actually appears to be his underlying position, explicitely endorsed in a number of articles. So, in 'Facts of the Matter' Quine clarifies his physicalistic understanding of factuality:

> Mental states and events do not reduce to behavior, nor are they explained by behavior. They are explained by neurology, when they are explained. But their behavioral adjuncts serve to specify them objectively. When we talk of mental states or events subject to behavioral criteria, we can rest assured that we are not bandying words; there is a physical fact of the matter, a fact ultimately of elementary physical states. (Quine, 1979a, p. 167)

Likewise, in the concluding remarks of 'Mind and Verbal Dispositions' Quine claims:

> Until we can aspire to actual physiological explanation of linguistic activity in physiological terms, the level at which to work is the middle one; that of dispositions to overt behavior. Its virtue is not that it affords causal explanations but that it is less likely than the mentalistic level to engender an illusion of being more explanatory than it is. (Quine, 1975c, p. 95)

Quine's comments suggest that scientific explanation in behavioural terms is merely a *temporary* substitute of a fully neurophysiological causal level of explanation. The claim, thus, that there is no fact of the matter with regard to what the terms of a language refer to, is not objectively guaranteed by the behavioural dispositions observed by the linguist, but rather by the physical states underlying those dispositions. My defence of the Inscrutability Thesis depends crucially upon favouring a full-blooded neurophysiological level of explanation. Thus, in chapters

---

[10] In chapter 7 we'll see that *replacement* of mental phenomena, rather than *reduction*, is the most

4, and 5 I shall make use of a connectionist (neurobiologically-inspired) model of cognition; see 4.4 below. The previous discussion seems to show that Quine could make himself at home in such a connectionist setting, transposing thus his old behaviouristic arguments for the Inscrutability Thesis into the neuroscientific fashion (see chapter 5, section 5.7 below).

On the other hand, the above discussion on levels of explanation has sometimes fostered the illusion that Quine's thesis is epistemological. Before closing this section, let me briefly address this issue, since it is vital to fully appraise the relevance of the Inscrutability Thesis upon the metaphysical status of semantics. Quine himself has not explicitely distinguished ontological and epistemological versions of his thesis.[11] Some commentators, however, focus upon what source of evidence should be relevant to translation, favouring thus an epistemological reading. The claim then is that the alleged plurality of translation manuals will always be compatible with the totality of acceptable *evidence*. The foe of Quine who targets this epistemological reading may then try to argue (even granting a neurophysiological level of explanation) that, were we to take into account all 'physically statable evidence', there would be a fact of the matter as to what the reference of a given term is.

However, under this epistemological reading, the attack would be harmless to Quine. Quine's focus on a behaviouristic level of explanation reflects the point that behavioural facts are the only ones relevant to fixing the semantic notion of reference. As I mentioned earlier, plausibly, more facts are relevant to semantics—

---

likely outcome once we endorse a neurophysiological level of explanation.

[11] See Friedman (1975) for an analysis of the distinction between ontological and epistemological versions of the Inscrutability Thesis.

e.g., neurosciencific facts. But to exploit the existence of more relevant evidence against Quine would be to miss the target. Thinking that we could dissolve the inscrutability of reference by checking whether a given neurophyiological state, A, (and not B) is the one actually causing mental state A* (and not B*), is missing the very point the Inscrutability Thesis is aiming to show. Namely, that it is meaningless to maintain that a speaker actually believes A* and not B*, or *vice versa.* The inscrutability of reference is not a claim about hidden neurophysiologically determined semantic facts. It is not a matter of lack of information about the architecture of the brain. The Inscrutability Thesis claims that even if we had all the relevant information about all the elementary physical mechanisms responsible for cognition, the choice would remain unsolved. The reason, as Quine notes, is simply that we cannot expect to find 'a distinctive mechanism for every purported distinction that can be phrased in traditional mentalistic language' (Quine, 1970d, p. 180). In conclusion, the discussion about which facts determine or fail to determine the semantic facts should not lead the reader into thinking that Quine's thesis rests upon an epistemological foundation. Its strength must reside in its ontological version, to the effect that the totality of acceptable *facts*—not evidence—fails to determine reference.

Enough of preliminaries. In the next two sections I shall recount Quine's parable of Radical Translation as aiming to support the Inscrutability of Terms. My exposition, for ease of explanation, will be faithful to Quine's original behaviouristic setting. Only later (chapters 4, and 5), I shall re-state the thesis in connectionist terms, and try to show its full potential in a connectionist, neurophysiologically inspired, reading.

## 1.3 *First Steps in Radical Translation: Stimulus Meaning and Observation Sentences*

In his well-known parable of *Radical Translation*,[12] Quine brings into play an ideal situation in which there is no connection whatsoever between the speakers of two different languages. One is the Native language; the other, the linguist's under which the inquiry will take place. The task is to reconstruct the Native language by means of a translation manual. This manual, when finally completed, should be able to correlate each of the potentially infinite number of sentences uttered by the natives with one or more sentences belonging to the linguist's Home language. The linguist is not allowed to correlate Native expressions with those of the Home language on the grounds that they pin down the same *idea*. Quine's naturalism, as we saw earlier, forbids us to visit the *Museum of Ideas* (see 1.1 above). Unable to pair words with language-independent mental acts, the linguist starts from scratch, acknowledging as a genuine evidential basis only the stimulation of her sensory receptors. Upon this she will try theories in search of true prediction. The process is the following.

---

[12] Due to the enormous amount of literature in the last four decades on Quine's project of Radical Translation, I shall try to go very briefly in this opening chapter over the details of the parable. The reader interested in the fine-grained detail is urged to visit the *locus classici*: Quine (1960), chapter 2; and (1969a), chapter 1. The reader interested exclusively in the Inscrutability Thesis may consult Quine (1969a); whereas, for interesting links between that thesis and the thesis of Indeterminacy of Translation, the place to go is Quine (1960). For a very good critical analysis of Quine's Inscrutability Thesis, and links to other well-known theses of Quine, the reader may care to consult Kirk, 1986. See also Wright (1997).

The linguist starts from an articulated web of Native sentences with no assumptions as to how they are going to be analytically dissected into their constitutive terms (see below).[13] The articulated web of sentences is supplemented by a behaviourally based distinction between *occasion* sentences, and *standing* sentences. The Native sentences that the linguist can start checking more easily are occasion sentences. Quine defines occasion sentences as those sentences that command assent or dissent only if queried after an appropriate current stimulation—see Quine (1960), pp. 35-6. Occasion sentences are the ones native speakers initially acquire in their own language, and constitute the entering wedge into Native for the linguist. The linguist, however, must make use of any semantic notion, defined in terms of stimulations, that helps her correlate occasion sentences of Native with occasion sentences of Home. To achieve this, Quine makes use of the behavioural notion of *stimulus meaning*:

> The stimulus meaning of a sentence for a subject sums up his dispositions to assent to or dissent from the sentence in response to present stimulation. (Quine, 1960, p. 34)

By putting together the set of all those stimulations that would prompt the native's assent with the set of all those that would prompt her dissent, the linguist can make use of a naturalistic notion of equivalence. Occasion sentences of Native and Home

---

[13] This already represents an advantage for the linguist. Namely, that she can start to devise her manual of translation regardless of what objects singular terms refer to. In 'Five Milestones of Empiricism', Quine reminds us of a salutary maxim introduced by Bentham. It reflects the idea that the minimum significative semantic unit is no longer the *word*, but rather the *sentence* as a whole—see Quine (1981b), pp. 68-70.

get correlated in virtue of having the same stimulus meaning—i.e., in virtue of being *stimulus synonymous*.

The set of occasion sentences of a language, nonetheless, does not constitute a monolithic block. Their membership is rather a matter of degree. Some will be easier to identify, some others more difficult. In this way, Quine distinguishes a priviledged subset of occasion sentences for the project of Radical Translation. Namely, those sentences whose stimulus meaning varies least under the influence of collateral information (cf. Quine, 1960, p. 42). Quine dubs them *observation* sentences. Their comparative constancy of stimulus meaning renders them suitable to form the basis on which the linguist will build up her translation manual. The linguist will be able to correlate an observation sentence, n, of Native with another one, h, of her Home language by noticing that the native assents to/dissents from n in every occasion in which the linguist would have assented to/dissented from h. Observation teaches the linguist to discriminate between occasion, and in particular, observation sentences on the one hand, and standing sentences on the other. From a range of situations the linguist can then try observation sentences upon the natives in different environmental situations, and inductively construct a tentative manual of translation. I shall not enlarge on the first steps of the linguist's translation manual (see Quine, 1960, pp. 26-30). Let us move on to see how the linguist's behaviouristic setting, aided by the notion of stimulus meaning, bears onto reference.

Quine considers an utterance of the native one-word observation sentence 'Gavagai'. By observing that on all the occasions in which the native had assented to/dissented from 'Gavagai?', the linguist would have given the same response to

'There is a rabbit?', the linguist inductively arrives to the conclusion that 'Gavagai' is stimulus synonymous with the English sentence 'There is a rabbit'. Concerned with stimulus meaning, the key point is that the linguist understands that the native sentence is a *rabbit-related* one. So far, no reason to worry about terms; the linguist's entry to Native is via sentences, and in particular those highly linked to present observable events. So, for the sake of the argument, let's assume without further ado that via the linguist's aforementioned inductive process 'Gavagai' is correctly translatable as 'There is a rabbit'.[14]

Bigger worries arise when we shift our attention from sentences to terms. As we've seen, observation sentences, and usually one-word ones, are fairly easy to pin down. But, unluckily, not all sentences are like this. Standing sentences cannot be directly correlated with some current stimulation. They will be assented to/dissented from irrespectively of the amount of perceptual similarity that both the native's and the linguist's sensory receptors might share at the time of the query. The linguist, therefore, is unable to translate *whole* Native standing sentences into Home sentences. The only way for the linguist to accomplish her task, Quine points out, is by making use of *analytical hypotheses*:

> [The linguist] segments heard utterances into conveniently short recurrent parts, and thus compiles a list of native 'words'. Various of these he hypothetically equates to English words and phrases. (Quine, 1960, p. 68)

---

[14] The reader should notice that the whole process depends on assuming that the linguist can distinguish native's assent from dissent (see Quine, 1960, pp.29-30). I shall ignore the possibility that the linguist is mistaken about what she takes to be native linguistic signs for assent and dissent. However, for some skeptical comments on this issue see Levy (1970), pp. 598-9.

In devising analytical hypotheses, the linguist shifts her attention from sentences (initially taken as the minimum semantic unit) to terms. The way the linguist dissects standing sentences is guided by the observation sentences already written down in her notebook. If a segment of a standing sentence appears in as an observation one, the linguist will start by matching other examples of this segment in other observation sentences of Native. The matching should reflect the previous pairing between the observation sentences. In this way, the linguist will be tempted to conclude that, for instance, the native term 'gavagai' can be equated with our Home term 'rabbit'.

However, according to Quine, we need not assume that the native term 'gavagai' refers to the set of rabbits on the basis that the Native and English related sentential counterparts are stimulus synonymous. The reason is simply that we could still retain the identity of stimulus meaning of 'Gavagai', and 'There is a rabbit', while arguing that 'gavagai' as a term divides its reference over things other than rabbits. Well-known putative examples are undetached parts of rabbits, their temporal stages, or any other perverse referent that 'gavagai' might divide its reference over which does not violate Quine's behavioural adequacy conditions (see chapter 3 below). Furthermore, the native term could be equated with some Home expressions that do not divide their reference at all. These are cases such as the abstract singular term 'rabbithood', standing for the universal; or 'rabbitfusion', that denotes the scattered region of space-time composed of all rabbits. Likewise, 'gavagai' could also be translated as the feature-placer 'rabbiting', which would stand in analogy with, for instance, 'raining' or 'snowing'. Bearing in mind Quine's behavioural setting, we can see why stimulus synonymy at the sentential level is not

affected by such semantically perverse proposals:

> Point to a rabbit and you have pointed to a stage of a rabbit, to an integral part of a rabbit, to the rabbit fusion, and to where rabbithood is manifested. Point to an integral part of a rabbit and you have pointed again to the remaining four sorts of things; and so on around. (Quine, 1960, pp. 52-3)

Thus, all the available evidence being the linguistic, and non-linguistic dispositions of the native speakers under observable circumstances, it cannot be confirmed that 'gavagai' refers to the set of rabbits as opposed to any of the aforementioned perverse alternatives. In short, we could be in possession of more than one correct manual of translation. All of which would agree which Native and Home sentences should be ascribed identical stimulus meaning. However such manuals are mutually incompatible since the terms of Home correlated with a given word of Native by each manual pick out different sets of objects in the world. In this way, we may have a standard manual that equates the native term 'gavagai' with 'rabbit', whereas a perverse manual might equate the same native term with, let's say, 'undetached rabbit part'. In conclusion, there is no objective fact of the matter as to which manual is the correct one, and what the extension of the Native term 'gavagai' is.

## 1.4 *The Juggling Strategy*

Quine's original argument in favour of the plurality of schemes of reference is not completed yet. What has been shown so far is the lack of relevant criteria in order to fix the reference of 'gavagai' as a term. But someone might think that the 'gavagai' example, as it stands, is not significant. The reason for this has nothing to do with

the way the argument has been developed, but rather with the example that has been chosen to exemplify the thesis. One-word observation sentences, such as 'gavagai', are a extreme case. The whole sentence takes the form of a single noun; it is not accompanied by any other grammatical devices such as singular or plural endings, definite or indefinite articles, counting expressions, etc. Someone might, thus, object that the plurality of choice will disappear as soon as we pay attention to more complex constructions in which 'Gavagai' is no longer a one-word sentence, but has rather been inserted into a bigger one. These worries drive us to the last part of Quine's project of Radical Translation.

The kind of sentential constructions we should pay attention to are the ones involving, for example, *Identity* as in '... the same as ...', *Quantity* as in 'there are x ...', *Plurality*, etc. [15] The anti-Quinean argues that by paying attention to the apparatus of individuation, perverse alternatives à la Quine will be behaviourally discredited. Hence, by asking the native, let's say, whether *there are two or three* so-and-so present, we may be able to tell whether the so-and-so is a term of divided reference or a mass term, for instance.[16] If the native is able to answer the question, that could count as evidence in support of the thesis that the so-and-so divides its reference. On the other hand, abstention of judgment may count as evidence in support of the thesis that 'gavagai' refers to objects not subject to such division.

---

[15] For the reader not familiar, these are the sort of constructions Quine refers to as the *apparatus of reference*—see Quine (1973), esp. Part III. In this chapter, and chapter 3 below, I shall refer to this sort of constructions as the *apparatus of individuation.*

[16] Notice that this strategy already involves a crucial assumption that I shall grant for argument's sake. Namely, that the linguists are able to ask questions in Native where the apparatus of individuation is present.

Imagine then, for the sake of the argument, that we could ask the native whether certain gavagai is or is not the same as the one she saw the day before, or about the number of gavagai present at a the time of the query. In this way (borrowing, and recasting an example from Hookway, 1988, pp. 148-9) if the linguist asked the native: 'Cuántos gavagai hay allí?', she may translate the native's answer (let us say, 'Dos gavagai') as 'There are two rabbits', and she would feel confident enough about her translation manual because of what she observes about the native's environment and her linguistic behaviour. However , as Quine points out, it may have been rash of the linguist to reject the other putative alternative translations of 'gavagai':

> We could equate a native expression with any of the disparate English terms 'rabbit', 'rabbit stage', 'undetached rabbit part', etc., and still, by compensatorily juggling the translation of numerical identity and associated particles, preserve conformity to stimulus meanings of occasion sentences. (Quine, 1960, p. 54)

Let us illustrate how this juggling would work: When the linguist translates standardly 'Dos gavagai' as 'There are two rabbits', she is pressumably employing the following principle of translation:

**(a)** 'dos' => 'there are two' and
**(b)** 'gavagai' => 'rabbits'.[17]

---

[17] The reader should notice that '=>' here is not a logical device. So far we are only deploying translation rules in a loose sense. I use '=>' simply to reflect the fact that the terms appearing at both sides of the equation enjoy a similar role in their respective languages, such that stimulus synonymy at the sentential level is preserved. A more rigourous framework will be required when we move from manuals of translation to theories of semantics (sections 1.6, and 1.7 below).

The Quinean, however, does not need to do a very difficult adjustment in order to translate perversely 'gavagai' as 'undetached rabbit part'. By changing (a) for

**(a)*** 'dos' => 'there are two animals which are composed of',

then she will be able to replace (b) by

**(b)*** 'gavagai' => 'undetached rabbit parts'.

And now, by putting (a)* and (b)* together, the perverse linguist can make use of an alternative translation manual that renders the native utterance 'Dos gavagai' as 'There are two animals which are composed of undetached rabbit parts'.[18] Hence, by means of compensatory adjustments, Quine claims, the perverse manual is as compatible with the behavioural facts as the standard manual is assumed to be.

In like vein, we can compensatorily adjust all the rest of the Native expressions, and argue, for example, that the translation of the Native sentence 'Cuántos gavagai hay allí?', is not our standard 'How many rabbits are there over there?', but rather the perverse 'Of how many animals are there undetached rabbit parts over there?' (cf. Wright, 1997). Nonetheless, it is worth remarking that Quine acknowledges that we should employ the standard manual, instead of the perverse alternatives. The very point that the Quinean would like to stress is that that choice

---

[18] As the careful reader will have noticed, the perverse manual is obliged to specify that the undetached rabbit parts belong to two different animals. Otherwise, if we said, for instance, 'There are two undetached rabbit parts', we might be referring to two different parts of the same animal (see chapter 3, section 3.4 below).

is based completely upon *pragmatic* interests: No particular manual is *actually true*, against the others (see 1.3 above).

The reader familiar with the 'gavagai' literature will surely recall one problem with the above proposal that has been highlighted by Hookway. As Hookway (1988, pp. 149-51) notes, thanks to the juggling strategy, the perverse manual seems to cope satisfactorily with *gavagai-related* sentences. But what would happen if we were confronted with a sentence of Native, let's say, 'Dos rosas', that the standard manual translates correctly as 'There are two roses'? How could the perverse manual preserve stimulus synonymy? If the perverse linguist claims that 'rosas' must be translated as 'undetached rose parts', then she is obliged to do some adjustments elsewhere. Unfortunately, according to her manual 'dos' has been translated as in (a)* above. So, the perverse linguist would come out with something like 'There are two animals which are composed of undetached rose parts' as the translation of the Native utterance 'Dos rosas'. For obvious reasons, a perverse manual that deploys that translation would not be faithful to the evidence. It seems then that Quine's juggling strategy fails when we start dealing with things other than rabbits.

However, on behalf of Quine, Hookway offers a solution to the problem. The perverse linguist could produce a somewhat more cumbersome manual by changing (a)* above for the following *disjunctive* rule of translation:

**(a)**** 'dos' => 'two animals which are composed of', when dealing with rabbit-related utterances,

or

'dos' => 'two plants which are composed of', when dealing with rose-related ones.

So, with the help of a disjunctive rule of translation the perverse manual can cope both with rabbits and roses. Although it is not difficult to guess the next move of the anti-Quinean. How would the perverse manual translate a new Native sentence that the standard manual has matched correctly with, for instance, our 'There are two stones'? The solution would be to insert another disjunct in (a)** in order to account for mineral-related utterances. And, now the anti-Quinean can do the same move once again, and so on, and so forth. The result is that the perverse linguist would come out with a translation manual which conforms to the evidence but which is extremely cumbersome. Whether such *ad hoc* manual can still be taken to be as correct as the standard one is something that I shall leave unanswered until chapter 2.[19]

This completes my introductory review of Quine's parable of Radical Translation.[20] In the remainder of this chapter I shall address two important counter-

---

[19] The reader familiar with the literature will have noticed that (a)** is actually different from Hookway's original version. His is hybrid in the sense that he employs a perverse disjunct for counting rabbits and a standard one when dealing with any other sort of objects. I find it more realistic to go for a fully-perverse manual all the way down. However, whatever choice we make (fully perverse, or standard-cum-perverse à la Hookway) will not influence my overall purposes (see chapters 2, and 3 below).

[20] We have now the appropriate background to shed some light upon my opening remarks in section 1.1. As I said earlier, Quine's parable of Radical Translation supports the thesis of referential inscrutability, rather than the thesis of indeterminacy of translation. The following passage from Quine reveals the reason for this: "The *gavagai* example was at best an example only of the inscrutability of terms, not of the indeterminacy of translation of sentences. As sentence, *Gavagai* had a translation that was unique to within stimulus synonymy; for the occasion sentences 'Rabbit', 'Rabbit stage', and 'Undetached rabbit part' are stimulus-synonymous and holophrastically

examples to the Inscrutability Thesis that Evans (1975) has produced. On behalf of the Quinean, I shall offer a rejoinder to Evans by expanding on Hookway's above disjunctive strategy. But before that, let me just make some closing remarks that will help clarify the importance of the Inscrutability Thesis. The lack of any sort of ideological or cultural resemblance among the users of Native and English certainly helps to make Quine's point more vivid. It is noteworthy, however, that even if the Inscrutability of Reference were true with respect to speakers of Native, its real significance would be missed unless we transfer the parable of Radical Translation to *home*. Someone might object to Quine's parable that bilingualism may help to solve the indeterminacy. The linguist could go to the native tribe, live with them and learn their language in the same way a native child would do. In short, if the linguist became bilingual, someone may argue, she would be able to discover which is the correct manual. This, nevertheles, should not cause any concern. As Quine points out,

> [it] makes no real difference that the linguist will turn bilingual and come to think as the natives do—whatever that means. For the arbitrariness of reading our objectifications into the heathen speech reflects not so much the inscrutability of the heathen mind, as that *there is nothing to scrute.* (Quine, 1969a, p. 5)

---

interchangeable. The *gavagai* example had only this indirect bearing on indeterminacy of translation of sentences: one could imagine with some plausibility that some lengthy nonobservational sentences containing *gavagai* could be found which would go into English in materially different ways according as *gavagai* was equated with one or another of the terms 'rabbit', 'rabbit stage', etc. This whole effort was aimed not at proof but at helping the reader to reconcile the indeterminacy of translation imaginatively with the concrete reality of radical translation." (Quine, 1970d, p. 182).

So, the problem is not that the linguist could force her own conceptual repertoire of Home into Native when learning the native tongue. We are assuming that the linguist does not take anything for granted. The point is that there is no singular individuation machinery to be assumed or not at home. There is not *one* conceptual repertoire belonging to Home. When two fellow speakers match their utterances by the *homophonic rule* (that is, by translating each other expressions phoneme by phoneme), their situation is not different from the one in Quine's parable. There decision not to employ a *heterophonic* manual is due exclusively to reasons of simplicity. For purposes of communication, smoothness is important; but in terms of fidelity of speech to evidence, heterophonic and homophonic transcriptions are on a par. In short, the ontological commitments of the assertions of two speakers of the same language—even of the same speaker at different times—are inscrutable. They can be interpreted as picking rival referential relations in the world. Let us now turn our attention to Evans' criticism of the Inscrutability Thesis.

## 1.5 *Locating Schemes of Predication*

Evans (1975) produced a line of argument which suggests that semantically perverse translation manuals à la Quine are *behaviourally* incorrect. In my opinion, the interest of Evans' argument relies in the fact that, unlike some foes of Quine that insist in the need of honouring a mentalistic level of explanation (see 1.2. above), Evans' attack is launched from within a Quinean framework. Evans tries to show that the perverse referential schemes will not be able to cope with all the data that the standard scheme does. And Evans confines himself to a pool of data which

Quine would acknowledge as genuine evidential basis: Namely, native assent and dissent to the linguist's queries under concurrent observable circumstances. If Evans' attack is sound, it may prove fatal since Quine will not be able to reply by claiming that Evans' criticism relies on non-factual considerations. Evans' anti-Quinean line of argument is a powerful one, and I shall spend some time in this, and the following section to review it.[21] A full appraisal will be crucial since a major thread of my defence of the Inscrutability Thesis arises as a reaction to Evans' attack (see chapter 3 below).

Evans starts by pointing out the divergencies between the task of a translator and the task of a semanticist. The aim of the former is simply to facilitate communication between two linguistic communities. In order to do so, she must devise a manual of translation. Evans does not manifest any concern with the claim that translation suffers from indeterminacy. The reason is simply that a translator is not devoted to revealing any *semantic* truth. The translator's aim is simply to find smooth vehicles of communication, and insofar as this target is achieved, the way the translator dissects native utterances (i.e., what analytical hypotheses she projects into Native—see section 1.3) is completely irrelevant to her task. By contrast, the semanticist is involved in the project of constructing a theory of meaning. She is not concerned merely with correlating expressions of Native with lumps of Home language, but rather with stating what the native expressions actually *mean*.[22] The

---

[21] Many philosophers take Evans' counter-examples to have definitely defeated the thesis of referential inscrutability—see, for instance, Kirk (1986), p. 47.

[22] In fact, Evans' approach differs from the original project of Radical Translation in more substantial respects. Being concerned with semantics, we need the concepts of truth, denotation, etc. And Evans' approach to such notions must be understood in a full-blooded sense: "[The] semanticist

sentences of Native are potentially infinite in number. The semanticist, similarly to the translator (see 1.3), will be obliged to dissect native sentences. The target now, however, is to account for the meaning of those previously unencountered native utterances in a recursive way. But in opposition to the case of Radical Translation, Evans claims, not any given set of analytical hypotheses will do. Quine's treatment of certain compound expressions, as we shall see next, is the root of Evans' distrust.

Quine claims that mastering compound observation sentences ('White rabbit', 'Yellow paper', etc.) is on a par with mastering simple observation sentences ('Rabbit', 'Paper', etc.). Speakers ostensively learn the use of bigger observational constructions in the familiar inductive way in which the use of one-word observation sentences is learned (Quine 1974, pp. 59-60). This similarity at the sentential level carries over to the theory of reference when we move down to the level of terms. Learning an observational term is learning when to assent to/dissent from it as an observational sentence. Since the learning of both simple and compound expressions follow the same pattern, the result is that the same referential indeterminacy that afflicts terms in simple observation sentences (see 1.3 above) afflicts also terms in compound sentences. The compound 'white rabbit' (as a term) is subject to the Inscrutability Thesis in the same way as the term 'rabbit' is meant to be. Even though 'white rabbit' relates to a portion of space-time, in the vicinity of the speaker, which is both *rabbit-related* and *white-related*, it would be rash to impute our ontology to the speaker. Quine maintains that the extension of the second component, 'rabbit', could be taken to be the set of undetached rabbit

---

aims to uncover a structure in the language that mirrors the competence speakers of the language have actually acquired." (Evans, 1975, pp. 343-4). In chapter 2, we shall see how Evans tries to exploit this issue to his advantage. See also fn. 25 below.

parts.[23] The conclusion is that it is indeterminate what the compound 'white rabbit' refers to: The semanticist may assign to the compound as its extension either a subclass of the set of rabbits or a subclass of the set of undetached rabbit parts. Our only hope, in Quine's view, of solving the indeterminacy is by looking at the interaction of such expressions with the *apparatus of individuation* (plurals, identity, etc.). Unfortunately, as we saw in section 1.4, this hope is thwarted, since the apparatus of individuation is itself inscrutable too.

Evans disagrees with Quine's contention, and notes that for Quine's argument to work,

> it must rest upon the belief that the sole reason a semanticist can have for treating an expression as a predicate with a particular divided reference is to account for that expression's interaction with the (putative) apparatus of individuation. (Evans, 1975, pp. 345-6).

However, according to Evans, the apparatus of individuation is not the only way to identify an expression as a predicate that refers to such-and-such objects. As a matter of fact, the location of the scheme of predication takes place at a subtler level. And such a prior anchoring determines the apparatus of individuation. The apparatus of individuation can, thus, only be secondary. Evans advances what this subtler level consists of:

> The primary function of construing an expression, G, as a term dividing its reference over rabbits—or trees—is to explain how the truth conditions of

---

[23] In like vein, the Quinean may produce a perverse rendering of the first component, 'white' (Quine 1974, pp. 81-3). However, we can ignore this additional perversity, since Evans is exclusively concerned with the semantic treatment of 'gavagai' (Evans 1975, p. 363).

> certain elementary, but compound, sentences into which it enters are determined by their parts. The apparatus of individuation may be entirely absent from such sentences. To see the notion "what G is a predicate of" in this way is to see it as constrained by a theory of sentence composition into which it fits and which alone gives it sense. (*Ibid.*, p. 346).

There is in particular one phase in the route to linguistic competence where Evans starts to dig in to launch his attack. This is the process by which two observational expressions are attributively combined as occurs for example in 'White rabbit' or 'Yellow paper'. Attributive composition can be better appraised by contrasting it with a more primitive form of combining one-word observation sentences. That is *conjunction*—see Quine (1990), p. 4 and (1970b), pp. 9-10. This mode of combination differs from the attributive one in the fact that it does not require any sort of overlapping of the features referred to by the two conjuncts. In this way, 'White here and rabbit here?' will be assented to whenever there is both a rabbit and *something* white in the speaker's vicinity. By contrast, the attributive compound 'White rabbit?' will only be assented to when a *decent-sized portion* of the rabbit is itself white. The overlapping of both features is the crucial point that distinguishes what Quine calls *mereological functions* (as in the case of attributive compounds) from truth functions (as, for example, conjunction). But, as Evans remarks, mereological functions are not enough for predication. What we do when we take an expression as a mode of predication is "to associate with it a certain condition, upon whose satisfaction by objects depends the truth or falsity of the sentences in which the expression occurs". (Evans, 1975, pp. 348-9).

Mereological functions are assented to when there is a significant overlapping of the features compounding the function. But, clearly, it is not sufficient for the

speaker to be confronted with a decent-sized portion of *rabbit* which is white. Think for example of several non-white rabbits with white tails which are distributed in such a way that there tails form a continuous white picture—cf. Evans, 1975, p. 351. In that case, speakers would still dissent from 'White rabbit?' This illustrates the fact that mereological functions cannot be taken as predicative constructions. In short, 'White rabbit' (taken as a mereological function) does not amount to 'There is a white rabbit here'. What we require to obtain the correct assent conditions, and this is the key point, is that the white feature gets distributed in a characteristic way in relation to the boundaries of a single rabbit, such that its presence prompts assent to the query 'White rabbit?' (cf. Evans, 1975, p. 351).

Hence we may say that 'White rabbit' is true if, and only if, there is a rabbit in the vicinity of the speaker which satisfies the condition of being white. In this way, a first step in individuation has been achieved. By identifying the construction 'White rabbit' with 'There is a white rabbit here', the truth conditions of such compound sentences can be easily explained. The contribution of the parts ('white' and 'rabbit') can be explained by taking 'rabbit' to be associated with a particular divided referent (i.e., a whole enduring rabbit), and 'white' as being distributed within the boundaries of such a particular object in a certain homogeneous way. So, summing up, the position Evans has arrived at is the following:

> To say that an expression has a particular divided reference makes sense only in the context of the explanation of compound sentences. To decide that a term divides its reference over rabbits is to decide that the sentences into which it occurs involve predication of rabbits. And to decide that a set of sentences involve predication of rabbits is to identify the way those sentences' assent conditions are generated from their parts as depending upon the

> identity conditions of rabbits, and so systematic mastery of those sentences requires mastery of the identity conditions of rabbits. (Evans, 1975, p. 355).

The Quinean, however, need to disagree with Evans in order to preserve her perverse alternatives. Explaining the aforementioned compounds as involving predication of rabbits may help the semanticist to build her theory up. Nevertheless, no reason has been given to support the view that the same role cannot be played by appealing to different ways in which the compound terms might divide their reference. For instance, over temporal stages of whole enduring objects, or over their undetached parts. Evans is aware of this point, and his aim in the last part of his seminal paper 'Identity and Predication' is to provide evidence to show why alternative perverse referents cannot possibly deliver the goods to the Quinean. Evans develops several different counter-examples in order to disprove one by one the various Quinean *ad hoc* alternatives (see 1.3). In what follows, I shall focus exclusively on Evans' treatment of 'undetached rabbit part'.[24]

---

[24] It must be stressed that even if the forthcoming arguments that I shall be offering against Evans' first counterexample—see section 1.6, and chapters 2, and 3 below—were on the right track, that would not have a straightforward bearing upon Evans' (1975) other counterexamples (for a thorough appraisal of these other counters, the reader may care to consult Wright, 1997). However, it must be noted that for the Inscrutability Thesis to work, the sympathiser of Quine may simply stick to *one*

## 1.6 *Evans' First Counter-Example, and Hookway's 'Divide-and-Rule' Strategy*

To introduce Evans' first counter-example, consider two *semantic theories* of Native, one standard and the other perverse.[25] One of the native expressions is 'Blanco gavagai'. Natives utter 'Blanco gavagai' *only* when a white rabbit shows up in their visual field. On the one hand the Standard Theory, ST, deals with 'Blanco gavagai' in the following way:

**ST**

Axioms:

**(a)** (x)(x satisfies 'gavagai' iff x is a rabbit)

**($a_1$)** (x)(x satisfies 'blanco'^$f$ iff (x is white & x satisfies $f$))

Theorem:

---

perverse alternative to standard semantic theorizing. In the remainder of this work, I shall thus restrict my interest to Evans' treatment of 'undetached rabbit part'.

[25] Although noted in section 1.5 that Evans' approach differs from Quine's insofar as the semanticist, unlike the translator, is concerned with semantic notions, the Quinean need not disagree. A sympathiser of Quine can concede that referential inscrutability actually concerns indeterminacy in the semantic field. By transferring Quine's original formulation into semantics, we fear no loss: Any theory of semantics will have ot match Native with Home sentences. And in doing so the semanticist relies upon the same body of evidence as the translator does. Namely, native assent to/dissent from queries under concurrent observable circumstances. From now on then I shall follow Evans, and illustrate Quine's perversity by means of theories of reference, rather than translation manuals.

**($a_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is a rabbit))[26]

On the other hand the alternative offered by the perverse semanticist is:

**$PT_1$**

Axioms:

**(b)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**($b_1$)** (x)(x satisfies 'blanco'^*f* iff (x is white & x satisfies *f*))

Theorem:

**($b_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is an undetached rabbit part))

Let us suppose that ST is behaviourally adequate. We can, thus, identify the sentence 'Blanco gavagai' with 'There is a white rabbit'. However, Evans argues, if ST is behaviourally adequate, then $PT_1$ is not behaviourally adequate. There are certain circumstances in which $PT_1$ fails to reflect correctly the native's linguistic behaviour (Evans 1975, p. 358)—assuming ST does correctly reflect the native's linguistic behaviour. The sort of situation Evans is thinking of is for example when native speakers are stimulated by a brown rabbit with a white leg. In this case, $PT_1$ is not faithful to the evidence since, assuming $PT_1$, natives should assent to 'Blanco gavagai?' when stimulated by a white-legged brown rabbit.[27] But, we have assumed

---

[26] ($a_2$) is obviously a consequence of (a) and ($a_1$). The reader might be expecting that 'theorems' of the standard theory would assign truth to sentences. However, it is simpler to stay with satisfaction for nothing in my ensuing argument hangs on the difference.

[27] Notice that a brown rabbit's white leg is a white undetached rabbit part.

that ST is behaviourally correct, and hence that natives would assent to the combined construction 'Blanco gavagai?' only in presence of a white rabbit.

There is a further alternative that Evans himself advances. In order to avoid the inconvenient consequences of white-legged brown rabbits, the obvious move is to link the satisfaction conditions of 'blanco' to things which are parts of white rabbits. The perverse theory would then require an axiom of the form:

**($b_1$)*** (x)(x satisfies 'blanco' iff x is an undetached part of a white *rabbit*)

But this move only brings further difficulties: What will the native say about white sheets of paper, snowed landscapes, and so on? It seems that we are obliged to extend the scope of ($b_1$)* in order to talk about white things other than rabbits. Hence, the broader axiom required should run as follows:

**($b_1$)**** (x)(x satisfies 'blanco' iff x is an undetached part of a white *thing*)

But unfortunately, as Evans notices, the Quinean still faces a similar worry to the one motivated by white-legged brown rabbits. According to ($b_1$)**, 'Blanco gavagai?' should be assented to when a claw of a white-legged brown rabbit is present. For the claw itself is a part of a white thing: namely, a white leg. At this point, Evans doesn't pursue these matters further. It seems there is nothing the Quinean can do.

Hookway, however, proposes a rejoinder to the difficulties which Evans has raised for the Quinean thus far. He contends that the problem arising with ($b_1$)* does not force us to go for ($b_1$)**. If we want to refer to white sheets of paper or snowed landscapes, then the way to do so is by displaying the satisfaction

conditions of 'blanco' in a *context-sensitive* way.[28] In order to do so Hookway (1988, p. 155) offers the following disjunctive axiom:

**$(b_1)$***** (x)(x satisfies W if, and only if, either

(a) W occurs 'together with' H and x is a part of a white animal

or

(b) W occurs in some other context and x is white).

Recasting Hookway's axiom $(b_1)$*** in our terminology we get:

**$PT_2$**

Axioms:

**(c)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**$(c_1)$** (x)(x satisfies 'blanco' iff either

(a) 'blanco' occurs together with 'gavagai' and x is an undetached part of a white animal

or

(b) 'blanco' occurs in some other context and x is white)

Theorem:

**$(c_2)$** (x)(x satisfies 'blanco'^'gavagai' iff (x is an undetached part of a white animal))

Hence, if the native utters 'Blanco gavagai' we employ the first disjunct of $(c_1)$. Otherwise, we use the second.

---

[28] Hookway's move is similar to the way in which he adjusts the apparatus of individuation to favour a particular perverse scheme via disjunctive rules of translation (see 1.4).

The careful reader may have spotted a difficulty with Hookway's proposal that prevents $PT_2$ from preserving its empirical adequacy. To wit: According to the theorem generated by $PT_2$, ($c_2$), native speakers should assent to 'Blanco gavagai?' in the vicinity of a white cat or a white cow. The reason is obvious. Notice that any undetached part of a white cat or a white cow is an undetached part of a white *animal*. I ignore what moved Hookway to formulate his proposal in terms of animals. However, it should not cause great inconvenience, for the modification required is minimal. By substituting 'rabbit' for 'animal' in the first disjunct of ($c_1$), we shall obtain the correct satisfaction theorem. Hence the perverse semantic theory Hookway requires is:

**$PT_3$**

Axioms:

**(d)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**($d_1$)** (x)(x satisfies 'blanco' iff either

(a) 'blanco' occurs together with 'gavagai' and x is an undetached part of a white *rabbit*

or

(b) 'blanco' occurs in some other context and x is white)

Theorem:

**($d_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is an undetached part of a white *rabbit*))

Now, $PT_3$ is behaviourally correct if the standard theory, ST, is—as required. Natives will *only* assent to 'Blanco gavagai?' in presence of a white *rabbit*. When dealing with white cats or white cows the second disjunct, (b), of ($d_1$) will come to

the rescue.[29] Hookway's disjunctive strategy, as reformulated in $PT_3$, seems to succeed in eluding Evans' counter-example. Unfortunately for the Quinean, as we'll see next, Evans is not ready to surrender yet.

## 1.7 *Evans' Second Counter-Example: Widening the Scope of Hookway's Disjunctive Route*

Evans (1975) is not specially worried about potential rejoinders to his first counter-example (1.6 above). The reason is that even if a positive solution to the 'white-legged brown rabbit' counter could be given, as Hookway's strategy suggests, there are further problems that Evans thinks the Quinean will not be able to face. Evans switches to another line of attack which he believes is lethal for the Quinean. As Evans suggests, if the native language under study contains

> some unstructured expressions whose satisfaction conditions were given on [ST] by use of 'is partly red' and 'is partly green', then, when we permute, it might be thought that these expressions could be given conditions satisfied by some but not all parts of the same rabbit. Thus the sentence whose truth conditions used to be given by ['There is a rabbit here and it is partly red'] is now rendered by the use of ['There is a rabbit part here and it is red'] —which has incontestably the same stimulus meaning [...] This theory, however, will not work if [ST] did. For the sentence ['There is a rabbit here and it is partly red and it is partly green'] would occasionally elicit assent, inexplicable upon

---

[29] The reader might expect the second disjunct in axiom ($d_1$), (b), to behave perversely too, as (a) does, keeping thus with the Quinean spirit. I address this issue in an introductory caveat in chapter 2, section 2.1, below. For the purposes of this chapter it suffices to appraise how Hookway's 'divide-and-rule' strategy manages to elude Evans' objection by means of a context-sensitive axiomatic base.

> the new theory, under which the sentence emerges as true iff there is a rabbit part present which is both red and green. (Evans, 1975, pp. 359-60).

Evans contends that the putative empirical adequacy that Quinean perverse semantics enjoys will be lost as soon as we pay attention to more complex Native sentences. Any perverse counterpart of the standard English sentence 'There is a rabbit here and it is partly red and it is partly green' would fail to preserve stimulus synonymy, misrepresenting thus Native usage. I don't think that Evans' second counter-example can sink Quine's project. In the remainder of this section I shall offer an extension of Hookway's proposal which avoids loosing its empirical adequacy (assuming Evans' hypothesized data). According to Evans, an extension of our standard theory, ST, would cope with, say, the *unstructured* Native expressions 'parcial-rojo' and 'parcial-verde' in the following way:

**ST$^+$**

Axioms:

**(a)** (x)(x satisfies 'gavagai' iff x is a rabbit)

**($a_3$)** (x)(x satisfies 'parcial-rojo'^$f$ iff (x is partly red & x satisfies $f$))

**($a_4$)** (x)(x satisfies 'parcial-verde'^$f$ iff (x is partly green & x satisfies $f$))

Theorems:

**($a_5$)** (x)(x satisfies 'parcial-rojo'^'gavagai' iff (x is partly red & x is a rabbit))

**($a_6$)** (x)(x satisfies 'parcial-verde'^'gavagai' iff (x is partly green & x is a rabbit))

**($a_7$)** (x)(x satisfies 'parcial-rojo'^'parcial-verde'^'gavagai' iff (x is partly red & x is partly green & x is a rabbit))

Assuming that $ST^+$ is behaviourally adequate, natives would assent for example to the combined construction 'Parcial-rojo, parcial-verde gavagai?' only in presence of a rabbit which is partly red and partly green. Now the challenge for the Quinean is to produce an alternative semantic theory to ST. A sympathiser of Hookway may contend that his disjunctive strategy can be deployed once more in order to cope with the hypothesized pool of data. Hence, our perverse semantic theory $PT_3$ (see 1.6 above) might be thought to deliver the goods via the following extension:

**$PT_3^+$**

Axioms:

**(d)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**($d_3$)** (x)(x satisfies 'parcial-rojo'^□ iff either

(a) □ = 'gavagai' and x is an undetached part of a partly red rabbit,

or

(b) □ = *f*^'gavagai' and (x is an undetached part of a red rabbit and x satisfies *f*^'gavagai'),

or

(c) 'parcial-rojo' occurs in some other context and x is partly red)

**($d_4$)** (x)(x satisfies 'parcial-verde'^□ iff either

(a) □ = 'gavagai' and x is an undetached part of a partly green rabbit,

or

(b) □ = *f*^'gavagai' and (x is an undetached part of a green rabbit and x satisfies *f*^'gavagai'),

or

(c) 'parcial-verde' occurs in some other context and x is partly green)

Theorems:

**($d_5$)** (x)(x satisfies 'parcial-rojo'^'gavagai' iff (x is an undetached part of a partly red rabbit))

**($d_6$)** (x)(x satisfies 'parcial-verde'^'gavagai' iff (x is an undetached part of a partly green rabbit))

**($d_7$)** (x)(x satisfies 'parcial-rojo'^'parcial-verde'^'gavagai' iff (x is an undetached part of a partly red, partly green rabbit))

The reader can see that by linking the satisfaction conditions of 'parcial-rojo'/'parcial-verde' (when concatenated with 'gavagai') to things which are parts of partly red/partly green rabbits, we can avoid the results Evans predicted for the Quinean. The reason is simply that the partly-*f* feature applies to whole rabbits (as is the case under ST), rather than to their parts. Hence, according to $PT_3^+$, native speakers will assent to the query 'Parcial-rojo, parcial-verde gavagai?' in exactly the contexts hypothesized by Evans: Namely, when a partly red, partly green rabbit appears in their visual field. Stimulus synonymy, *contra* Evans, is thus preserved.

Someone, however, may maintain that the proposal does not go far enough. Put bluntly, the problem for $PT_3^+$ stems from the fact that the native expressions 'parcial-rojo' and 'parcial-verde' are not actually *unstructured.* Plausibly, by observing the linguistic behaviour of native speakers we'll soon realise that they can talk of a wide variety of objects in their environment as being partly *f* (where '*f*' need not stand for a colour feature, but rather for any other property natives might ascribe to the object in question). So, they will be able to say, for instance, that a bottle is *partly* empty.[30] But in that case, any semantic theory aiming to explain the

[30] There is no point in arguing that, unlike English speakers, natives might lack the apparatus to construct this sort of combined expressions. As I mentioned earlier, Quine's Inscrutability Thesis is meant to apply at Home (see 1.4 above), where we do employ 'partly' in many different contexts.

linguistic competency of native speakers must tell us what the semantic value of 'parcial' is. By taking into account the systematicity manifested in the natives' behaviour, we shall be able to determine the semantic value of 'parcial *f*' out of the satisfaction conditions of the semantically primitive expressions 'parcial' and '*f*'. In this way, we'll be able to grasp the semantic contribution that the simple term 'parcial' makes in 'parcial rojo' as well as in for example 'parcial vacío' (the expression that ST has correctly equated with our 'partly empty').

Nevertheless, acknowledging that 'parcial rojo' and 'parcial verde' are structured expressions should not occasion any distress to those sympathetic to Hookway's route—at least insofar as we manage to produce a semantic theory containing separate axioms for 'parcial', 'rojo' and 'verde'. The perverse semantic theory thus required, $PT_3^{++}$, would include the following disjunctive axioms:[31]

**$PT_3^{++}$**

Axioms:

**(e)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**($e_1$)** (x)(x satisfies 'rojo' iff x is red)

**($e_2$)** (x)(x satisfies 'verde' iff x is green)

**($e_3$)** (x)(x satisfies *f*^'gavagai' iff (∃y)(x is an undetached part of y & y is a rabbit & y satisfies *f*)

**($e_4$)** (x)(x satisfies 'parcial'-*f*^□ iff either

[31] I leave as an exercise for the reader to modify the standard theory, ST, (see section 1.6) so that it can account for the satisfaction conditions of 'parcial', 'rojo' and 'verde' in a *structured* way.

(a) ▯ = 'gavagai' and (∃y)(x is an undetached part of y & y is a rabbit & (∃z)(z is an undetached part of y & z satisfies *f*), [32]

or

(b) ▯ = *g*^'gavagai' and (∃y)(x is an undetached part of y & y is a rabbit & y satisfies *g* & (∃z)(z is an undetached part of y & z satisfies *f*)

or

(c) ▯ = 'parcial'-*g*^'gavagai' and (∃y)(x is an undetached part of y & y is a rabbit & (∃z)(z is an undetached part of y & z satisfies *f*) & (∃w)(w is an undetached part of y & w satisfies *g*)

or

(d) 'parcial-rojo' occurs in some other context and x is partly red)

The above axiomatic structure seems to deliver the goods to the Quinean. It delivers satisfaction theorems similar to those generated by $PT_3^+$—($d_5$), ($d_6$) and ($d_7$), so that the property of 'being partly *f*' applies to whole rabbits, rather than to their parts. Otherwise, Evans' counter would incontestably kick in. Thanks to $PT_3^{++}$, we have no problem to account for the different possible uses of 'parcial'. When natives talk about rabbits as being partly red, or partly green, we employ the first disjunct, (a) of ($e_4$), together with the appropriate supplementary axiom for the particular value of *f* in question. Hence we get, for example, the theorem

**($e_5$)** (x)(x satisfies 'parcial'^'rojo'^'gavagai' iff (∃y)(x is an undetached part of y & y is a rabbit) & (∃z)(z is an undetached part of y & z is red))

---

[32] It could either be the case that x=z or x≠z. Notice that all (a) in ($e_4$) is saying is that x is an undetached part of a rabbit which has a part that is *f*: It does not need to be the case that x is the part that the *f* feature applies to.

and also, for example

**($e_6$)** (x)(x satisfies 'parcial'^'rojo'^'parcial'^'verde'^'gavagai' iff (∃y)(x is an undetached part of y & y is a rabbit) & (∃z)(z is an undetached part of y & z is red) & (∃w)(w is an undetached part of y & w is green))

So, it seems that $PT_3^{++}$ can deal perfectly with, say, 'parcial rojo, parcial verde gavagai' when 'parcial rojo' and 'parcial verde' are taken to be structured. The satisfaction theorem for 'Parcial rojo parcial verde gavagai' is derivable from the semantic properties of its constituents. That is, from the semantic values of 'parcial', 'rojo', 'verde', and 'gavagai', specified respectively in axioms ($e_4$), ($e_1$), ($e_2$) and (e). So, the way ($e_6$) has been generated is the following. Since 'parcial rojo' occurs together with 'parcial verde gavagai', we employed the third disjunct, (c), of ($e_4$). Then in order to cash out the semantic vocabulary remaining in right hand side of (c) —i.e., z *satisfies* 'rojo' and w satisfies 'verde'—we employed axioms ($e_1$) and ($e_2$). As we can see, a native guided by $PT_3^{++}$ will assent to/dissent from the query 'Parcial rojo, parcial verde gavagai?' in exactly the same kind of contexts in which a native guided by ST would: Namely, when a rabbit which is partly red and partly green passess by. In conclusion, $PT_3^{++}$ is behaviourally adequate, and once again, stimulus synonymy, *contra* Evans, is preserved.

## 1.8 *Conclusion*

The discussion in sections 1.6, and 1.7 indicates that the Quinean need not be

embarked on a lost cause (at least, with respect to Evans' hypothesized data). We saw how Hookway modified Quine's perverse manual to make it behaviourally correct. Initially, Hookway's proposal failed to deliver the right satisfaction conditions. But, as I argued, Hooway's 'divide-and-rule' strategy could be expanded to bypass Evans' two counter-examples. Unfortunately, the problems for the Quinean are far from over. Hookway's route succeeded in its task at the expense of deploying a somewhat *barroque* axiomatic structure. In the next chapter I shall review a couple of criticisms due to Evans, and Wright that exploit the structural complexity of Hookway's route in order to discredit perverse semantic theorizing. The only way out for the Quinean seems to be to prove that structural simplicity cannot become alethic for semantic theories. That will be the topic for discussion of chapter 2.

# 2

## *IS SIMPLICITY ALETHIC FOR SEMANTIC THEORIES?*

### 2.1 *Introduction*

In chapter 1 we saw how an expanded version of Hookway's disjunctive strategy bypasses two counter-examples Evans (1975) offered against Quine's Inscrutability Thesis. Nevertheless, some philosophers have urged, Quine's thesis is not saved by Hookway's disjunctive reading. Evans (1981) and, more recently, Wright (1997) have argued on different grounds that, under certain conditions, *structural* simplicity may become alethic—i.e., truth-conducive—for semantic theories. Being structurally more complex than the standard semantic theory, Hookway's disjunctive route (see section 1.6, above) is an easy prey for Evans' and Wright's considerations. The bulk of this chapter will be devoted to addressing Evans' and Wright's criticisms. I shall argue that both Evans' and Wright's criticisms are unmotivated, and do not jeopardize Hookway's overall enterprise. But before that, let me just highlight a potential problem for Hookway that the careful reader may

have spotted in chapter 1.

Hookway is careful not to offer his perverse disjunctive proposal as conclusive against Evans, but rather as 'no more than a first approximation to a satisfactory response' (Hookway, 1988, p. 155). Hookway acknowledges the possibility that 'the attempt to develop [his] proposal consistently would run into technical difficulties' (*ibid.*, p. 155). Hookway makes these remarks with an eye to Evans' potential attack based on structural simplicity (see 2.2 below). However, there is a more basic technical hurdle for Hookway's proposal. Consider again how the Standard and Hookway's disjunctive semantic theories of Native dealt with the Native expression 'Blanco gavagai' (see chapter 1). Recall that natives utter 'Blanco gavagai' *only* when a white rabbit shows up in their visual field. On the one hand the Standard Theory, ST, deals with 'Blanco gavagai' in the following way:

**ST**

Axioms:

**(a)** (x)(x satisfies 'gavagai' iff x is a rabbit)

**($a_1$)** (x)(x satisfies 'blanco'^$f$ iff (x is white & x satisfies $f$))

Theorem:

**($a_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is a rabbit))

On the other hand, a version of Hookway's alternative to the standard route ST, as modified in chaper 1, runs as follows:

**$PT_3$**

Axioms:

**(d)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**($d_1$)** (x)(x satisfies 'blanco' iff either

(a) 'blanco' occurs together with 'gavagai' and x is an undetached part of a white rabbit

or

(b) 'blanco' occurs in some other context and x is white)

Theorem:

**($d_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is an undetached part of a white rabbit))

Although $PT_3$ is behaviourally adequate whenever ST is behaviourally adequate (see 1.6), the semantic perversity of $PT_3$ is rather narrow in scope. $PT_3$'s results coincide with the standard ones, as achieved via ST, except for rabbity expressions: The satisfaction conditions of 'blanco' are linked to *undetached parts* of white-... *only* when 'blanco' is coupled with 'gavagai'. In all other cases, $PT_3$ behaves standardly, taking 'blanco'-related utterances to be associated with whole enduring white cats or white sheets of paper, for example. This hybrid character of $PT_3$ (i.e., standard-cum-perverse) seems to be alien to Quine's original pursuit. Quine's aim was to produce a *fully perverse* alternative to ST in the sense that for *every* standard referent that ST picks out, a perverse counterpart is offered.[1]

Now, it seems that when we try to broaden the scope of Hookway's perverse route we are in trouble. If $PT_3$ is to account for Evans' first counter (see section 1.6) while being fully-perverse, ($d_1$) will have indefinitely many disjuncts. We will require an indefinite number of disjuncts in order to link the satisfaction conditions of 'blanco' to the appropriate wholes of undetached parts of rabbits, cats, cows,

[1] See Quine, 1973, esp. chapter 3.

paper, etc., etc. And the same will happen with respect to all those axioms required for dealing with any other Native colour-word, and indeed, with any other Native expression for which a version of Evans' counter can be put forward. Therefore, it *may* be the case that the perverse semanticist will not be able to state a fully-perverse *disjunctive* semantic theory.

However, in fairness to Hookway, we ought to notice that this difficulty is rooted on rather speculative grounds. First, it is unclear why the Quinean should not favour an array of merely hybrid semantic theories, rather than a single fully-perverse one. And second, even if the Quinean wishes to be fully-perverse, it is not obvious that the aforementioned difficulty could not be overcome by some baroque plot which the Quinean has up his sleeve (see chapter 3 below). Nevertheless, for the purposes of this chapter we need not expand on these considerations. Were we to concede the Quinean a position to fall back on for argument's sake, there are still two other criticisms due to Evans and Wright which, to many, seem to be crucial against Hookway. Let's take them in turn.

## 2.2 *Evans' Mirror Constraint*

In the closing passages of 'Identity and Predication' Evans remarks:

> I do not pretend to have shown that a viable semantic theory based upon one of Quine's suggestions cannot be constructed. Perhaps an ingenious person will show that the difficulties are less severe than they look, and thereby make something of Quine's *example* of the indeterminacy of semantics. (Evans, 1975, p. 363)

Despite this rethorical concession, Evans would not have been moved by

Hookway's strategy. According to Evans, a semantic theory, by contrast to a theory of translation, aims to provide a *psychological* explanation of the speakers' verbal behaviour by singling out certain behavioural dispositions (see 1.5). Speakers' behavioural dispositions, however, are to be understood in a full-blooded sense, thus marking a watershed with respect to the radical translator's account of disposition, where previously all that mattered was the preservation of stimulus meaning (see 1.3). In Evans' view, the semanticist must consider not only the linguistic regularities that natives may exhibit, but the underlying states that *explain* such regularities. In this way, we may talk not only of the behavioural dispositions themselves, but also of the *causal* explanatory states underlying those dispositions:

> I suggest that we construe the claim that someone tacitly knows a theory of meaning as ascribing to that person a set of dispositions—one corresponding to each of the expressions for which the theory provides a distinct axiom. [... It] is essential that the notion of a disposition used in these formulations be understood in a full-blooded sense. [... The ascription of tacit knowledge] involves the claim that there is a single state of the subject which figures in a causal explanation of why he reacts in this regular way to all the sentences containing the expression. (Evans, 1981, pp. 124-5)[2]

Once we conceive semantic theories as psychologically real—i.e., as tacitly endorsed by the speakers of a language—a *structural* constraint, alien to the radical translator's enterprise, comes into play. We should expect, Evans contends, that the derivational structure of the semantic theory tacily endorsed by the speakers of a language is somehow *mirrored* in the causal structures found in the speakers. That is the essence of what has become known as the *mirror constraint*. We may state it

---

[2] The reader may care to consult Davies, 1986, for a development of Evans' insights.

as follows:

**Mirror Constraint** The derivational structure of a semantic theory—i.e., the canonical route leading from the theory's axioms to the theorems produced—should mirror a causal structure found among the competencies of the speakers.

In a nutshell, the mirror constraint tells us that there must be an underlying causal explanation of the way competent speakers comprehend their language. And such a causal explanation will provide us with a picture of the actual route leading from the speaker's dispositions associated with the atomic elements of their language (its names and predicates, etc.) to the overall states associated with the whole sentences they might produce.

The mirror constraint proves to be a useful tool, for instance, when confronted with extensionally equivalent semantic theories. Given two theories that deliver the same set of well-formed theorems, we may decide, on empirical grounds, which is the correct one by looking at the dispositions of the speakers. To illustrate it, Evans (1981) introduces an example of an artificial language, L, consisting of 10 names ('a', 'b', 'c', ..., 'j') and 10 predicates ('F','G','H', ..., 'O'). A competent speaker of L should be able to produce, and understand when uttered by a different speaker, 100 different sentences—by coupling each name with all the predicates, one at a time. Nevertheless, we can be in possession of at least two semantic theories—call them $T_1$ and $T_2$—which are extensionally equivalent, agreeing thus in their specifications of the truth-conditions regarding the sentences of L. So, both $T_1$ and $T_2$ will specify, for example, that

'Fa' is true (in L) iff Pete is sad,

and so on for the rest of the sentences of L. But whereas $T_1$ is composed of 100 axioms—one for each well-formed sentence of L—, $T_2$ has only 21 axioms—10 for the names, 10 for the predicates and 1 for the coupling of a given name with any predicate. So the problem resides in choosing between $T_1$ (*listiform* theory) and $T_2$ (*articulated* theory).[3]

So far, the point we are interested in is how to use the mirror constraint in order to discover which is the correct theory. That is, can the causal structure underlying the speaker's dispositions justify us in choosing one theory, as opposed to the other? At this point, Evans appeals to different empirical evidence we may have in order to explain speaker's linguistic abilities, such as patterns of acquisiton or loss of understanding of sentences of L. We can observe what happens to the competencies of speakers when loosing competence of a particular sentence. Thus, imagine that as a result of a stroke a speaker looses competence with 'Fa'. If by loosing competence with 'Fa' we observe that the speaker's competence with the other 99 sentences remains intact, then we should infer that $T_1$ is the correct theory.[4] Whereas if, on the other hand, we observe that when a speaker looses competence with 'Fa' she looses as well competence of all those other sentences in which either 'F' or 'a' play a role, *and* she keeps intact her competence with all other sentences, *and furthermore*, something similar happens when the original sentence she lost her

---

[3] See Miller (1997) for a comprehensive review of the mirror constraint.

[4] Obviously, this follows from assuming that the choice is restricted to $T_1$ and $T_2$—i.e., to a *fully* listiform semantic theory, and a *fully* articulated one. Hybrid options would oblige the semanticist to withold his judgement until a wider range of evidence is considered.

competence with is not 'Fa' but any other sentence, then we shall say that we are justified, on empirical grounds, in choosing $T_2$. So, it seems that the mirror constraint provides us with a way to determine empirically which semantic theory is the correct one.[5]

Although Evans' target has been discrediting rival semantic theories which yield the same theorems as the standard one, it's not difficult to see how Evans's argument applies to cases where the theories under consideration are not thus equivalent, such as our standard theory, ST, of Native, and Hookway's disjunctive alternative, $PT_3$—see 2.1.

## 2.3 *Full-blooded Semantics and Disjunctive Semantic Theories*

One obvious advantage of Evans' mirror constraint with respect to Quines scepticism towards the Theory of Reference is that hopefully a unique choice of referential scheme will be *empirically grounded.* This is because the tacit knowledge ascribed to the native speakers comes in terms of the causal-explanatory states attributed to natives—or better said, to their internal information-processing systems. And the structure of these states will clearly not be discovered from the armchair. Once we know what this structure is like, what the mirror constraint tells us is that it will *mirror* the derivational structure of the theory that is implicitly known. In this way, the anti-Quinean may deploy this further *structural* constraint; a constraint which is alien to the radical translator's original enterprise, and which will thwart, Evans believes, any Quinean hope of transferring translational

[5] For present purposes we need not go into the detail of Evans' argument. The reader may care to consult Evans (1981), and Wright (1981) for a rejoinder to Evans.

indeterminacy to the semantic field.

Evans (1975) contends that those semantic theories that divide the reference of 'gavagai' over undetached parts of rabbits may have a chance of working. Nevertheless, that's only at the expense of attributing to the speakers of the language *unwarranted dispositions*—see Evans (1975), p. 363. Confronted with ST and $PT_3$, we may find that native speakers do follow, though tacitly, ST, and not $PT_3$, by observing for instance that mastering the term 'blanco' in contexts which do *not* include 'gavagai', permits natives to *understand* such expression in *all* contexts. If this were to be the case, then this behavioural evidence would favour ST over $PT_3$, since native speakers would have just *one single disposition* for judging sentences containing 'blanco' as having such-and-such truth-conditions—as opposed to having two different dispositions, as occurs under $PT_3$: one to account for the first disjunct, (a), of ($d_1$), the other for the second disjunct, (b). On the other hand, were we to observe that loosing competence with 'blanco' in *any* non-rabbit context left unaltered the native's understanding of 'blanco' when coupled with 'gavagai', that would count as partial evidence for $PT_3$.

Hookway (1988, pp. 155-62) considers Evans' view that one semantic theory may give a better psychological explanation of a speaker's verbal behaviour than another, but he believes that it poses no serious threat to $PT_3$. His reason is that the Quinean would simply reject as *non-factual* any psychologically-based criterion which goes beyond the description of the observable behaviour of speakers. Hookway's Quinean notes that

> unless psychological explanations simply allude to physical mechanisms, they do not enhance our knowledge of (physical) reality. (*ibid*, p. 159)

However, Hookway is overlooking a crucial point: Namely, that Evans' argument can be transposed into a form which a physicalist will have to admit as legitimate. The key point is that a physicalist should expect there to be some relation between speaker's linguistic manifestations and the information content of *physical states* in their brains, such that the canonical route in a theory of meaning leading from its axioms to the theorems produced reflects a *neurophysiological* causal structure found underlying the competencies of the speakers.[6] This means that there should be a neurophysiological explanation of the way competent speakers comprehend their language. And this causal explanation will provide us with a picture of the actual route leading from the speaker's dispositions associated with the atomic elements of Native to the overall physical states associated with the whole sentences they produce. Hopefully, just one semantic theory will thus be *empirically grounded* since the tacit knowledge of the semantic theory ascribed to a certain speaker of Native comes in terms of the causal explanatory states attributed to her internal information-processing system.

Once we know how this internal system operates, it is theoretically plausible that we can determine whether a speaker tacitly follows ST or $PT_3$. If future neuroscience reveals that there is one single neurophysiological state causally activated when a native utters 'blanco' in all different contexts, then that would

---

[6] This is indeed Evans' original approach: "The decisive way to decide which model is correct is by providing a causal, presumably neurophysiologically based, explanation of comprehension" (Evans 1981, p. 127). In Quine's view this is the correct level of analysis: "To cite a behavioural disposition is to posit an unexplained neural mechanism, and such posits should be made in the hope of their submitting some day to a physical explanation" (Quine 1975, p. 95)—see chapter 1.

count as evidence against $PT_3$, since assuming $PT_3$ we would require two different neurophysiological states: Namely, one state exclusively responsible for 'blanco' when coupled with 'gavagai' and a different one causally responsible for all other 'blanco'-related utterances.[7]

Being an open empirical question whether there is actually a body of evidence favouring ST over $PT_3$, let me close this section with two speculative remarks: Evans believes that considerations concerning the productivity and systematicity of language and thought, would tip the balance in favour of ST. Evans' well-known *generality constraint* claims that:

> if a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception. (Evans, 1982, p. 104)

By assuming the generality constraint we are committed to a demand for a causal systematicity in relation to our concept-mastering abilities.[8] If we are to explain a set of inferences by appealing to a common piece of concept mastery, then there *must* be some internal factor which is common to all the inferential transitions. In short, by acknowledging the generality constraint we are demanding a single inner state which gets activated whenever a cognitive episode involving a given concept

---

[7] Someone might object that speakers who follow $PT_3$ are after all being tacitly guided by a simple, though *compound*, dispositional state. Namely, the disposition to assent to 'blanco gavagai' when there is an undetached part of a white rabbit in the vicinity of the native, *and* to assent to 'blanco *f*' when there is a white *f*. The onus however would be on neuroscientists to explain what such a neurophysiological state would look like. This is nevertheless an open empirical question.

[8] We may read Evans' generality constraint as dealing with linguistic utterances, rather than concepts, for nothing hangs on the difference.

occurs.

Unfortunately, we still lack the neuroscientific apparatus to judge whether the generality constraint is correct or not. In chapter 4 I shall argue that *if* certain connectionist models of human cognition are on the right track, then Evans' mirror constraint and generality constraint have little chance of working.[9]

On the other hand, a second caveat concerns whether the mirror constraint does indeed favour ST over *any* perverse semantic theory the Quinean might produce. Someone may argue that the mirror constraint needs to be supplemented by some sort of *uniqueness* constraint, such that a system of dispositions will be empirically found to back ST, *and* that no more than one semantic theory will be correct under these empirical findings. Again, I must postpone development of this point until chapter 3, where I shall offer a perverse route which survives both the mirror constraint and the putative uniqueness constraint just canvassed.

Nevertheless, the anti-Quinean, without being committed to either the mirror constraint or the generality constraint, can still adduce further considerations aiming to discredit perverse semantic theories. Before I develop my main arguments in chapters 3 and 4 in defence of Quine's inscrutability thesis, we must look at a different argument based again on structural simplicity considerations which Wright has recently offered to show that Hookway's disjunctive strategy cannot be a viable alternative to ST.[10]

---

[9] The reason—to advance one of the main points of my research (see chapter 4, below)—is that we need not posit tacit rules in order to explain the productivity and systematicity of language and thought (see Elman, 1998).

[10] For an overview of Wright's appraisal of Quine's argument for the Thesis of the Inscrutability of Reference and the reasons he produces against it, see Wright (1997).

## 2.4 *Wright's "Metodological Simplicity" Criterion*

Crispin Wright (1992) has reshaped debates about Realism by offering a new landscape of what's at stake in the discussions between realists and their opponents. Instead of arguing whether a given discourse can be truth apt, discussion should focus, Wright contends, on what *kind* of truth predicate a discourse can enjoy. Namely, whether truth for a discourse can be 'robust' or merely 'minimal'. Wright's approach has important implications for Quine's Thesis of the Inscrutability of Reference. The remainder of this chapter will be devoted to showing that an argument involving minimalism about truth which Wright (1997) offers against the Inscrutability Thesis fails by *reductio*. By the end of the chapter, we'll see how Wright's proposed frame for discussion of Realism bears on the metaphysical status of Semantic Theories.

A difficulty is raised by Wright (1997) which, if accepted, would favour ST over $PT_3$—see 2.1. Wright introduces a criterion of 'methodological simplicity'.[11] Wright admits that in general simplicity is not alethic:

> [Simplicity] cannot be assumed, without further ado, to be an *alethic*—truth-conducive—virtue in empirical theory generally. There is prima facie sense in the idea that of two empirically adequate theories, it might be the more complex that is actually faithful to the reality which each seeks to

[11] By 'methodological simplicity' Wright refers to the sort of structural simplicity considerations which seem to weaken Hookway's disjunctive route. Note, however, that Wright would not agree with Evans' structural simplicity argument as reviewed in the last section for Wright wouldn't buy the mirror constraint—see Wright, 1981, and Miller, 1997.

> circumscribe. (*Ibid.*, p. 411)

Although not alethic in general, simplicity is alethic, Wright thinks, in certain circumstances. There must be some 'further ado' that will transform what is not initially alethic into something alethic.

> [The] thought that, when it comes to radical interpretation, there is an ulterior psychologico-semantical reality which an empirically adequate translation scheme might somehow misrepresent is, of course, exactly what Quine rejects—exactly what he famously stigmatizes as the myth of the semantic museum [see chapter 1]. And with that rejection in place, methodological virtues which are not, in realistically conceived theorizing, straightforwardly alethic can now become so. In such cases, the methodologically best theory ought to be reckoned true just on that account. It is therefore not enough for a defender of Quine to seek to save the alternative schemes by postulations which, though still principled and general, are comparatively expensive in terms of ambiguity and other forms of complication. If a simpler scheme is available, that fact is enough to determine that these alternatives are *untrue*, by the lights of the only notion of truth that, in Quine's own view, can engage the translational enterprise. (*Ibid.*, p. 411)

We can expand on the thinking behind Wright's remark if we look at his discussion in *Truth and Objectivity*. Wright interprets Quine as an anti-realist about meaning. Applied to the case of discourse about meaning, the discussion in *Truth and Objectivity* allows that discourse about meaning will be apt for minimal truth and falsity, and some semantic theory may well be true. It is sufficient to be fitted for minimal truth, Wright contends, that a discourse meets the constraint of *disciplined syntacticism*: (a) The discourse must have *sufficient* discipline to support a practice of warranted assertion (i.e., the use of sentences must be *standardly regulated* such

that uttering 'p' will or will not be considered appropriate depending of the situation in question). And, (b) the discourse must exhibit a number of syntactic possibilities that permits speakers to say things like 'if p&q then p', 'not-p', 'I believe that p', etc. Disciplined syntacticism will ensure, Wright holds, that the discourse in question does deal with *bona fide* assertoric contents apt for truth and falsity. However, discourse about meaning will fail certain other tests by which a discourse qualifies as realist, and hence does not qualify for a 'substantial' notion of correspondence between true sentences of the discourse and the facts that make them true. Thus, Wright thinks that where a substantial notion of correspondence with the facts is in play, simplicity is not alethic. The more complex of two equally epistemically justified theories may be the one which corresponds to the facts. But where a minimal notion of truth is in play, the simpler theory is the true theory.

The key test by which Wright distinguishes a realist discourse from a discourse apt for mere minimal truth is 'cognitive command'. A discourse exerts cognitive command if, and only if:

> It is a priori that differences of opinion formulated within the discourse, unless excusable as a result of vagueness in a disputed statement, or in the standards of acceptability, or variation in personal evidence thresholds, so to speak, will involve something which may properly be regarded as a cognitive shortcoming. (Wright, 1992, p.144)[12]

---

[12] Notice the importance of stating 'cognitive command' as an *a priori* constraint. We may find areas of discourse where it happens to be the case that no disagreement ever emerges. However, we should not conclude from that that the discourse in question achieves or lacks cognitive command. Whether a discourse exerts cognitive command or not must depend on the *content* of those expressions belonging to the discourse. For an elaboration of this idea see Wright, 1992, pp. 94, 168-70.

By 'cognitive shortcoming' Wright is thinking of any kind of shortcoming belonging to one of the following three broad categories:

> "divergent input", that is, the disputants' working on the basis of different information (and hence guilty of ignorance or error, depending on the status of that information), or "unsuitable conditions" (resulting in inattention or distraction and so in inferential error, or oversight of data and so on), or "malfunction" (for example, prejudicial assessment of data, upwards or downwards, or dogma, or failings in other categories already listed). (*Ibid.*, p. 93)

Wright's idea then is that cognitive command will help us to discern whether the true assertions of a discourse represent states of the world in a *genuinely* realistic fashion. If a discourse passes the test of cognitive command, then its true assertions may represent the states of the world in a heavyweight manner. So to speak, cognitive command may help to *beef up* the notion of representation. In contrast, when a discourse does not pass the test of cognitive command, the notion of representation to be applied to that discourse will be lightweight, though, we must remember, not weightless since minimal notions of truth and falsity apply.

## 2.5 *Beefing up Semantic Discourse: A Reductio contra Wright*

This brief review of Wright's approach to the realist/anti-realist debate will suffice to illuminate the bearing of Wright's argument upon the acceptability of Hookway's semantic theory $PT_3$. Take two contesting theorists who each support respectively one of the rival referential schemes ST and $PT_3$. They frame their respective theories in English. The question arises whether such theories, so framed, are apt

for truth and, if so, whether such truth is minimal or robust—all according to Wright's criterion. Wright is claiming that such theories are apt for minimal truth only, and that judged by the standards of minimal truth, only the standard theory is true. We may express Wright's argument as follows:

**Wright's version of Quine's assumptions**

**(i)** Truth for semantic discourse is minimal.
(Wright's reading of Quine's rejection of the Museum Myth).

**(ii)** Semantic discourse exerts assertoric discipline in that both the Standard Theory—ST—and its Quinean alternative—$PT_3$—are fully supported by the Native behavioural evidence.
(Assuming Quine's idealization about the facts of Native behaviour).

**Wright's premises**

**(iii)** If the truth for a discourse is minimal (and consequently the relation of correspondence and the facts which the discourse is about are minimal), then simplicity is alethic.

**(iv)** ST is simpler than its Quinean alternative—$PT_3$.

**Wright's conclusion from (i)-(iv)**

**(v)** ST is true and its Quinean alternative—$PT_3$—is false.

However, we may now continue the argument as follows:—

**Assumption**

**(vi)** ST and $PT_3$ are the only semantic theories which are fully supported by Native behavioural evidence.[13]

If we take premises (i)-(iv) and (vi) and add to them a further premise, then we may derive a contradiction from the whole set of premises:

**Additional premise**

**(vii)** If semantic discourse exerts cognitive command, then truth for semantic discourse is not minimal (and consequently the relation of correspondence and the facts to which the truths correspond are not minimal).

We proceed,

**(viii)** Semantic discourse exerts cognitive command.
(From (v) and (vi), since only the Standard Theory—ST—is assertable).

Therefore (*contra* (i)),

**(ix)** Truth for semantic discourse is not minimal.
(From (vii) and (viii)).

---

[13] Quine claims that there is an indefinite set of fully behaviourally adequate theories. For simplicity I have limited the alternatives to ST and $PT_3$, since this does not affect the substance of the argument. As is Wright, we are assuming ST is the simplest of all behaviourally adequate theories, and $PT_3$ is acting as a representative of all the rivals to ST.

We cannot accept the whole set of premises {(i), (ii), (iii), (iv), (vi) and (vii)} since they are inconsistent, yielding a contradiction between (i) and (ix). However, we shall not reject (i) or (ii) because we wish to stick, with Wright, to Quine's proposed frame of discussion. Nor will we question the simplifying assumption (vi). So, the premise to be abandoned must be either (iii), (iv) or (vii). In agreement with Wright, I shall not question (iv).[14] So, we are left with (iii) and (vii). I shall go for (iii)—arguing that simplicity cannot become alethic for semantic theories (even though we accept that the truth predicate to be applied to semantic discourse remains minimal). Wright's best shot then is to go for (vii)—the additional premise I assumed in order to obtain the *reductio*. So, let's see whether such an option is available to Wright.

Wright argues that passing the test of cognitive command is a *necessary* condition for a beefed up notion of representation—i.e., the kind of heavyweight representational status to be associated with a *robust* (not minimal) account of truth. However, Wright's position allows that cognitive command is not a sufficient condition for robust truth. He sums up his general position as follows:

> Suppose a class of predications such that it can never be a priori excluded that disagreement about one of them originates in some variation in a particular non-cognitive disposition of the disputants. Then there will be no obstacle to defining a range of concepts, cognate to those distinctive of the predications in question, such that nothing counts as a disagreement about the application of one of *these* concepts unless the disputants *share* the relevant non-cognitive disposition. By describing a disagreement as focused upon the

[14] At least for the purposes of this chapter. In chapter 3 I shall offer a perverse semantic theory in the line of $PT_3$ which is actually as simple, methodologically speaking, as the Standard Theory, ST.

> application of one of these latter concepts, we can thus preempt the possibility that it has one particular kind of non-cognitive source. *Nevertheless, the operation of the non-cognitive disposition is no less involved in the application of the new concepts than in that of the old. It is not that we now stand on firmer ground, or engage more "robust" matters.* (*Ibid.*, p. 224), (Last emphasis added).

An example of Wright's will illustrate the point. Suppose comic discourse—i.e., discourse about what is and is not funny—is sufficiently disciplined and has the right kind of syntax to support a truth predicate. Thus, practitioners of comic discourse will agree, for example, that it is not funny that I have two hands (in a 'normal' context). They will also agree that one who denied that Charlie Chaplin or Buster Keaton were funny would be deemed to be wrong and to lack a sense of humour. However, plausibly, comic discourse does not pass the test of cognitive command. Practitioners may disagree as to whether Buster Keaton is funnier than Charlie Chaplin, but agree that neither of them is in cognitive error. They may just have somewhat different senses of humour, such that their disagreement is not reducible to other areas of discourse concerning non-comic facts. The difference between their senses of humour is within the limits of what is normal for their community. Hence, following Wright, we are supposing that comic discourse does not exert cognitive command and is therefore apt only for minimal truth. Now, imagine a 'subcommunity of comic empathisers' (*Ibid.*, pp. 223-4) such that their senses of humour coincide perfectly. Whenever they are in disagreement about a comic situation, their disagreement is always explicable in terms of a dispute about non-comic aspects of the discourse. We could then imagine this subcommunity setting up a new discourse, comic*, which does exert cognitive command. Practitioners of comic* discourse introduce a new set of concepts such that it is *a*

*priori* that if two people disagree as to whether something is *funny**, then one of them must be in cognitive error. Hence, when comic empathisers disagree as to whether Buster Keaton is funnier* than Charlie Chaplin, they will maintain, in contrast to what we do in the case of comic discourse, that one or other is in cognitive error. This error involves a cognitive defect, a false appraisal about some non-comic* fact about the situation. However, Wright plausibly claims:—

> the mix of the cognitive and the affective in the basis for opinions about comedy* is exactly the same as it is for opinions about comedy. (*Ibid.*, p. 224)

Therefore, the fact that comic* discourse passes the test of cognitive command whereas comic discourse doesn't, does not imply that the first enjoys a beefed up notion of representation. It seems then that all they achieve by replacing concepts of the comic by concepts of the comic* is, as Wright puts it, "objectivity for cheap" (*Ibid.* p. 224). The moral of the comic* example is that cognitive command is *not sufficient* for realism. Thus, premise (vii) in the above derivation cannot be assumed without further ado.

However, the key issue, I shall contend, is that the results achieved in the area of comedy cannot be applied to the case of semantics. In general, I offer the following conjecture:

**(C)** If a discourse exerts cognitive command, then realism is the default presumption. The discourse is apt for robust truth unless reason can be found to downgrade the notion of truth involved in the discourse.

In the comic scenario reason was offered by Wright to show why comic* discourse is apt only for minimal truth, even though it passes the test of cognitive command.

As we've just seen, Wright argues that the *same non-cognitive disposition* is involved in both judgements—i.e., those judgements concerning the comic and those concerning the comic*. However, there is no parallel between comic and semantic discourses. By assuming that simplicity is alethic, semantic discourse exerts cognitive command. But there is no case for claiming that some non-cognitive disposition is involved in our semantic judgements. The dispute between an advocate of ST and an advocate of $PT_3$ doesn't depend on assuming vagueness in the Native statements under dispute or in the standards of acceptability, or on variations in personal evidence thresholds. Any disagreement, then, must involve a *cognitive* error on the part of the perverse semanticist since she is overlooking a crucial alethic datum: Namely, that the assertoric discipline of semantic discourse is subject to a principle of simplicity. And, simplicity being alethic, the perverse semanticist is guilty of 'malfunction'—a kind of cognitive shortcoming noted by Wright (*Ibid.*, p. 93). The perverse semanticist is *prejudicially assessing* the data, in the light of the availability of ST.

The position we have reached is this: Wright's best shot was to reject (vii),

**(vii)** If semantic discourse exerts cognitive command, then truth for semantic discourse is not minimal (and consequently the relation of correspondence and the facts to which the truths correspond are not minimal).

Wright has shown that in general cognitive command is not sufficient for robust truth, by the example of the comic empathisers. However, I have claimed that were we to take semantic discourse to pass cognitive command because simplicity is alethic, then there would be no reason to deny that semantic truth is robust. Hence,

denying (vii) does not give Wright a satisfactory way of avoiding the contradiction.

There is however a possible rejoinder. Wright might reject (C) by claiming that cognitive command and an additional condition are *jointly* sufficient for realism. This additional condition is that the discourse has an 'intuitional epistemology'. Wright connects the idea of robustness to the notion of an intuitional epistemology when discussing a hypothetical 'trivialising theorist' (*Ibid.*, pp. 148-57). The trivialiser complains that the test of cognitive command is vacuous. Given disagreement within a discourse, it is not clear whether the discourse satisfies the constraint since we don't know what to count as a *cognitive* shortcoming. If we take any disagreement within a discourse which intuitively fails the test of cognitive command, the trivialiser would claim that the disagreement actually involves a cognitive shortcoming since "ignorance or error will at least be involved *concerning the truth value of the disputed statement*" (*Ibid.*, p. 149). In order to avoid this risk of trivialisation, Wright brings into play the notion of an intuitional epistemology. That is, an epistemology such that our judgements concerning the subject-matter of the discourse are justified in a non-inferential manner. Wright contends that assertions that beefily represent the facts *need* to be backed by an intuitional epistemology.

Hence, Wright might reject conjecture (C) above. He may claim that truth for semantic discourse is minimal—even though it exerts cognitive command—because semantic discourse is not backed by an intuitional epistemology. We observe the verbal behaviour of the natives in their jungle setting. We do not observe semantic facts—see chapter 1. Our semantic discourse postulates semantic facts to explain the observed behaviour. Hence, it seems, Wright has found a reason why truth for semantic discourse is minimal even though it exerts cognitive

command.

I don't think that this line of argument will work. The reason is that it will not fit Wright's overall position. Wright would accept that, for instance, discourse about microphysical facts does exert cognitive command and is furthermore to be interpreted in a robust sense. Nonetheless, we lack an intuitional epistemology for microphysical facts, since we have intuitional access only to the observable facts of physics. Microphysics postulates unobservable—theoretical—facts beyond the reach of our observational capacities. And yet Wright would agree that these so called theoretical facts (i.e., the facts of microphysics) are robust. Parallel to this, it seems we have intuitional access to the behaviour of natives in their jungle context, and we postulate unobservable (theoretical) semantic facts to explain the verbal aspects of their behaviour. So, we can see that the lack of an intuitional epistemology does not provide us with a reason why semantic discourse exerts cognitive command and yet truth for that discourse remains minimal.

We saw that premises (i), (ii), (iii), (iv), (vi) and (vii) imply a contradiction. We are not questioning (i) and (ii) because they form the Quinean background to the discussion. We also granted (iv) and the simplifying premise (vi) for the purposes of this chapter. That left (iii) and (vii) as candidates for a *reductio*. We have now seen that (vii) stands. Hence, I conclude that it is (iii) which must go. Simplicity for semantic discourse cannot become alethic.

## 2.6 *The Metaphysical Status of Semantics*

Where does the preceding discussion leave us regarding the metaphysical status of semantics? We've seen that simplicity is not alethic. But this leaves us in an

uncomfortable position. It might seem now that semantic discourse has too little discipline to support even minimal truth. Wright makes some remarks about the amount of discipline required for minimal truth aptness in connection to comic discourse. For comic statements to be minimally true, we require a minimum of discipline such that practitioners of the discourse will agree that, for example, in a normal context it is not funny that I have two hands. The discourse may then become slack in other cases.

However, in the case of semantic discourse, taking simplicity not to be alethic, we find that there is *no* discipline at all. This may not seem obvious at first sight. The standard semantic theory, ST, and its perverse counterpart, $PT_3$, as spelt out in section 2.1, differ *only* on the satisfaction theorems that both theories generate respectively in order to deal with the Native sentence 'Blanco gavagai':

**ST:** **($a_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is a rabbit))

**$PT_3$:** **($d_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is an undetached part of a white rabbit))

But by looking exclusively to ($a_2$) and ($d_2$), someone might argue, we cannot justifiably claim that semantic discourse is not disciplined at all. We simply don't possess enough evidence to argue so. Nonetheless, by recalling Quine's approach to Radical Translation this worry disipates. According to Quine's setting of Radical Translation—see chapter 1—, the *only* agreement between standard and perverse translators is on logical constants, on signs of assent and dissent and on highly

observation sentences.[15] Quine's original pursuit was to produce a fully perverse alternative to ST in the sense that for every standard referent that ST picks out, a perverse counterpart is offered. Plausibly then, we may extend $PT_3$ such that for each theorem ($a_n$) of ST of the form:

**($a_n$)** (x) (x satisfies 'N' iff ... ),

the Quinean will produce a rival account that yields a theorem ($d_n$) of the form:

**($d_n$)** (x) (x satisfies 'N' iff ... )

where what fills out the dots in ($d_n$) differs in extension from what fills out the dots in ST's theorem ($a_n$). The idea is to achieve semantic perversity by producing a scheme of reference that conforms to all possible evidence, and yet assigns different extensions to *most* of the Native terms from those assigned by ST. This would ensure the aforementioned lack of discipline—insofar as ST and its perverse counterpart are behaviourally adequate and equally correct.

At this point the careful reader may have spotted a difficulty. As I pointed out in section 2.1, although being behaviourally adequate, the semantic perversity of Hookway's strategy is rather narrow in scope. The *hybrid* character of $PT_3$—standard-cum-perverse (see 2.1)—, someone may contend, might bring enough discipline for semantic discourse to support a truth predicate (at least, minimally).

---

[15] Though maybe not even this much could be given for granted. The disagreement might even be wider than initially expected. See Levy (1970) for scepticism about agreement on assent/dissent signs, and see Quine (1973), pp. 81-3, for a perverse rendering of highly observation sentences such as 'Red'.

In fairness to Wright, this may well be the case. Nevertheless, we need not dwell on this issue since in the next chapter I shall offer a perverse route which is *fully* perverse, undermining thus this potential problem for Hookway's *hybrid* alternative. Therefore, granting that the Quinean can produce a fully perverse semantic route—see chapter 3, below—I conclude that semantic discourse is not disciplined enough in order to enjoy the benefits of a truth predicate, not even the *benefits* of minimal truth.

This position is congenial to Quine—see chapter 1, and chapter 7, below. Quine is an eliminative materialist who claims there are no semantic facts. In contrast, if we make simplicity alethic in order to enforce sufficient discipline to support a truth predicate for semantic discourse, we get cognitive command and hence, as we saw, robust truth as well. We are thus jumping from no truth at all to robust truth. Wright has failed to show that there is metaphysical room for semantic facts which are not robust facts.

## 2.7 *Conclusion*

In this chapter I considered two arguments put forward by Evans (1981) and Wright (1997) respectively, which threatened Hookway's perverse semantic proposal. Exploiting the fact that Hookway's perverse semantic theory is structurally more complex than the standard theory, ST, Evans and Wright argued, on different grounds, that Hookway's route looses its empirical adequacy since, under certain conditions, structural simplicity may become alethic for semantic theories. The bulk of chapter 2 has been devoted to arguing that both Evans' and Wright's criticisms are unmotivated, and cannot jeopardize Hookway's overall enterprise.

Nevertheless, even if the arguments I've developed in this chapter were shown to be wrong, the Inscrutability Thesis would still not be endangered by Evans' and Wright's considerations. In the next chapter I shall pursue a perverse semantic route which differs substantially from the one advanced by Hookway. My proposal is as simple—structurally speaking—as its standard counterpart. Thanks to this feature, my strategy, I shall contend, is not subject to putative rejoinders along Evans' and Wright's 'structural-simplicity' lines. Furthermore, it is not subject either to certain other criticisms that I shall review in the next chapter; criticisms that tell against Hookway's proposal.

# 3

## *SEMANTIC PERVERSITY*

### 3.1 *Introduction*

In this chapter I shall propose a perverse theory of reference which differs substantially from the one advanced by Hookway in his attempt to bypass Evans' counter-examples (see chapter 1). In view of the results achieved in chapter 2, where I contended that Evans' and Wright's 'structural simplicity' considerations leave Hookway's proposal unaffected, let me outline the motivations for pursuing a different proposal. The perverse semantic translation manual I shall be offering is as simple, structurally speaking, as the standard translation manual, ST. Thanks to this feature, my strategy is not subject to certain criticisms which may put Hookway's proposal in jeopardy, thereby becoming an overall better candidate for the Quinean to fulfill her goal.

First, as I argued in chapter 2, Evans' 'structural simplicity' argument can be transposed into a physicalist format, threatening Hookway's disjunctive proposal.

Evans hoped to cut down the number of empirically adequate semantic theories to just one—namely, the standard theory, ST—by considering the neurophysiological states that speakers are being attributed as causally (explanatory) active, both in linguistic production and comprehension—see 2.3. I acknowledged that it is an open empirical question whether there is actually a body of neurophysiological evidence that favours those semantic theories that are structurally simpler. Whether or not Evans' mirror constraint—see 2.2—can deliver him the goods is a matter for future research in the neurosciences, and will ultimately depend on what kind of architecture embeds our higher cognitive abilities—see section 2.3, and chapter 4 below. But insofar as Evans' constraint is drawn from a physicalist framework, its bearing is a theoretical possibility that the Quinean cannot ignore. Fortunately for the Quinean, the proposal I shall advance, being as simple—structurally speaking—as ST, undermines Evans' considerations.

Moreover, my proposal has another advantage over Hookway's. In section 3.5, I shall consider an extension of Quine's succinct behavioural criteria of Radical Translation (see 1.2, and 1.3 above) suggested by Jaakko Hintikka's Game-Theoretical Semantics (1973; 1976). I shall argue that Hintikka's semantics suggest behavioural criteria which we can use to constrain perverse semantic theories. In particular, I shall try to show that whilst Hintikka's behavioural data tells against Hookway's disjunctive proposal, it reveals, nonetheless, further reasons (beyond structural simplicity) as to why my perverse semantic proposal enjoys the same privileged status that the standard theory, ST, is supposed to enjoy. So, without further ado, let's flesh out these considerations.

## 3.2 *A Perverse Way of Dividing Reference over Parts of Things*

Evans' (1975) attack on Quine's Inscrutability Thesis has been so widely well received by the philosophical community because of an implicit, though misleading, assumption made by foes and sympathizers of Quine alike. Namely, that reference is to be divided over objects in a *monolithic* fashion. Evans (1975, p. 362) talks in terms of semantic theories that cut the reference of 'gavagai' finer than the standard theory does—e.g., over undetached rabbit parts.[1] It is however tacitly assumed that finer cuts, such as the division of the reference of 'gavagai' over undetached rabbit parts, constitute a monolithic block. That is, the axioms that deal with the satisfaction conditions of 'gavagai' and 'gavagai'^*f* are spelt out such that *any* undetached rabbit part smaller than a whole enduring rabbit satisfies the argument.

However, I contend, we need not cluster *all* undetached rabbit parts under the same semantic theory. Rabbit claws, feet, legs and heads are undetached parts of rabbits. But we can differentiate among them, and articulate semantic theories whose axioms deal with those anatomical parts separately. In this way, 'gavagai', under one particular scheme, might be taken to divide its reference over undetached legs of rabbits, for instance; under another scheme, over undetached tails of rabbits; and so forth. Unfortunately, were the semanticist to specify *which* particular anatomical part of a rabbit her scheme makes use of, it would be fairly easy for the anti-Quinean to rebut the proposal. Simply by pointing; for even though every time you point to a rabbit, you are pointing to an undetached rabbit part, you need not

[1] For present purposes I shall ignore Quine's *coarser* cuts. The reader may care to consult Evans (1975). Wright (1997) offers a critical appraisal of all the different Quinean proposals (both finer and coarser) and of Evans' counters to all of them.

point to, say, its leg in every occasion. Therefore, the semanticist will be able to discard, on inductive grounds (see chapter 1) *a* particular undetached rabbit part as the target of the native's ostensive behaviour.[2] Nevertheless, there is a better option available to the Quinean.

I shall propose a particular way to discriminate among schemes of reference denoting diverse undetached rabbit parts that is not subject to the aforementioned difficulties. We may talk in terms of the percentage of the whole rabbit, including the percentage of its surface, that each scheme assigns as the extension of 'gavagai'. In this way, one putative scheme may claim that 'gavagai' divides its reference over 5% of the whole rabbit, including 5% of its surface (henceforth abbreviated 5%-urp: —i.e., 5% undetached rabbit part). Another scheme over 20%-urp; another over 80%-urp, and so on. Notice that neither *pointing* nor *questioning* (see chapter 1) can help to solve the referential indeterminacy. Every time you point to a rabbit, you are pointing to a 5%-urp, to a 20%-urp, to an 80%-urp, etc. Moreover, any further questioning beyond querying 'Gavagai?' that involves the apparatus of individuation (identity, plurals, etc.), will be dependent on imputing to the natives our ontology when interpreting such questions. By employing Quine's juggling strategy (1.5), we may take natives' assent to/dissent from any given query as evidence in favour of a 'x%-urp' scheme, as opposed to the standard one—see below.

---

[2] In fairness to Quine it must be noted that the case is not quite settled. It is unclear that the Quinean could not reestablish the empirical adequacy of her perverse scheme by means of some cumbersome plot which she has up her sleeve. However, the burden of proof is on the Quinean to make her case, and I fail to see how she could preserve structural simplicity, but I shall not press on this point.

Let's see how some semantic theories that cut the reference of 'gavagai' over x%-urp can cope with Evans' white-legged brown rabbit. Take, for instance, a perverse semantic theory that divides the reference of 'gavagai' over 5%-urp. Such a theory would include the following axioms:

**(p*)** (x) (x satisfies 'gavagai' iff x is a 5%-urp), and

**(p**)** (x) (x satisfies 'blanco'^*f* iff (x is white & x satisfies *f*))

Hence, taking the satisfaction conditions for 'blanco' in the standard way[3], our putative semantic theory will generate theorem (p):

**(p)** (x) (x satisfies 'blanco'^'gavagai' iff (x is white and x is a 5%-urp))

However, such a perverse semantic theory would not resist Evans' attack. A version of Evans' first counter-example (see 1.3) would kick in. Think of a brown rabbit which, instead of having a white leg, has 5% of its surface white-coloured. In this case, natives guided by (p) would assent to 'Blanco gavagai?' when stimulated by a 5%-white-coloured brown rabbit. Whiteness distributed all over a 5%-urp would not work since it elicits the wrong answer under certain circumstances. Semanticists agreed that natives would not assent to 'Blanco gavagai?' unless they are in presence of a white rabbit—see 1.5. And clearly an object which only has 5% of its surface Ø-coloured does not count as a Ø-coloured object.

The careful reader may have guessed by now what the next move for the Quinean should be. Evans' initial contention (see 1.5) about compound expressions

---

[3] Note that (p**) coincides with ($a_1$)—i.e., the axiom employed by the standard theory, ST—see 1.6.

such as 'blanco gavagai' was that 'blanco' had to be distributed in a particular way with respect to the boundaries of the object prompting native's assent to the query 'Gavagai?' The key word is *distribution.* In natural languages, when we say that a rabbit is white, we are assuming that the white feature is distributed more or less uniformly over all the surface of the rabbit. Let's say that when the percentage of white-coloured surface is equal or bigger than ß, then we take the rabbit as white.[4] Now, my contention is that a perverse scheme that divides the reference of 'gavagai' over ß%-urp will cope with Evans' white-legged brown rabbit. Take ß for instance as 99%. The perverse theory would then run as follows:

**$PT_4$**

Axioms:

**(e)** (x) (x satisfies 'gavagai' iff x is a 99%-urp)

**($e_1$)** (x) (x satisfies 'blanco'^$f$ iff (x is white & x satisfies $f$))

Theorem:

**($e_2$)** (x) (x satisfies 'blanco'^'gavagai' iff (x is white & x is a 99%-urp))

Now, let's see how this perverse referential scheme behaves under Evans' pool of data. The question is: Would the native guided by $PT_4$ assent to 'Blanco gavagai?' when a brown rabbit with a white leg is in his presence? Certainly not, for the native will only assent to the query when the 99% of the surface of the rabbit is

---

[4] I can set up the example in terms of percentage-of-*surface* (rather than volume) since we are restricting our attention to highly observational features such as 'colour' which applies to the external surface of objects. Notice, however, that since the 'x%-urp' scheme was defined in terms of x% of *whole* objects, including x% of their surfaces, we could bypass putative versions of Evans' counter that exploited volume features—like mass.

white. Hence, Evans' first counter-example is not a counter to $PT_4$. Those sympathetic to Evans would have to develop a different version of his counter in which the white portion of the brown rabbit is bigger. But not any bigger portion will do. We require the brown rabbit to have a white part occupying the 99% of its surface. But in this case, we would be confronted with a white rabbit, rather than with a brown one. Therefore, Evans' example is unable to show that $PT_4$ misrepresents native usage. A translator guided by this perverse scheme will predict native assent to/dissent from 'Blanco gavagai?' in exactly the same sort of situations in which a 'non-perverse' translator would. The reason is that rabbits and 99%-urp are *observationally indistinguishable*.

The reader can see that the '99%-urp' scheme differs from ST in a non-trivial way. What we need to achieve semantic perversity is a scheme of reference that conforms to all possible evidence, and yet assigns *different* extensions to the native terms from those assigned by ST. The following is *a priori*:

$(x)(y)\ (x = y \rightarrow (z)\ (z \text{ is a part of } x \leftrightarrow z \text{ is a part of } y)$

This condition establishes the semantic perversity of $PT_4$. Since 99 is smaller than 100, there will always be an undetached part of a whole rabbit which does not belong to the given 99%-urp: —namely, a 1%-urp. Hence, the perversity of $PT_4$ is *real* in the sense that the set of objects satisfying the property of being white does not coincide with the set of objects contemplated under ST.

Bearing in mind these results, we can now go back to one of the caveats left unanswered in chapter 2. In section 2.3, I called into question whether Evans' mirror constraint does indeed favour ST over *any* perverse semantic theory the Quinean might produce—granting for argument's sake that future neuroscientific

research might tip the balance in Evans' favour. As I hinted, someone may argue that the mirror constraint needs to be supplemented by some sort of *uniqueness* constraint.[5] The motivation behind this putative constraint was to secure that *no more than one* semantic theory will be correct under the potential empirical findings that the mirror constraint hopes to exploit. By comparing the semantic structure of $PT_4$ with that of ST, we'll soon realize that the Quinean has no reason to worry. Recall ST and $PT_4$:

**ST**

Axioms:

**(a)** (x)(x satisfies 'gavagai' iff x is a rabbit)

**($a_1$)** (x)(x satisfies 'blanco'^$f$ iff (x is white & x satisfies $f$))

Theorem:

**($a_2$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is a rabbit))

**$PT_4$**

Axioms:

**(e)** (x) (x satisfies 'gavagai' iff x is a 99%-urp)

**($e_1$)** (x) (x satisfies 'blanco'^$f$ iff (x is white & x satisfies $f$))

---

[5] For current purposes we need not worry about how to flesh out this additional constraint. The reason for this will become apparent in a moment. Nevertheless, note that the issue hinges on what *source* of evidence the uniqueness constraint can exploit. And at this point, the debate has been framed so that only behavioural and neurophysiological data is relevant (see section 1.2, above). Appealing to normative considerations to state a uniqueness constraint would beg the question.

Theorem:

**($e_2$)** (x) (x satisfies 'blanco'^'gavagai' iff (x is white & x is a 99%-urp))

Notice that derivations in $PT_4$ have exactly the same syntactic structure as derivations in the standard theory, ST. Therefore, if the data Evans hypothesized (see 2.3) showed that any semantic theory aiming to explain Native linguistic behaviour ought to do so by means of non-disjunctive axioms, $PT_4$ would conform to such a constraint. Evans suggested that by observing for instance that mastering the term 'blanco' in contexts which do not include 'gavagai' permits natives to grasp such expression in all contexts, that would count as evidence in favour of ST, as opposed to baroque alternatives such as Hookway's. However, $PT_4$ being as simple structurally speaking as ST, the chances for Evans to articulate such a constraint become slimmer. Natives who follow $PT_4$ will not be attributed unwarranted dispositions of the kind Evans suggests since natives would have just *one* single disposition for judging sentences containing 'blanco' as having such-and-such truth conditions. The indeterminacy, thus, remains unsolved. We haven't got a clue as to whether 'gavagai' divides its reference over rabbits or 99%-urp.

## 3.3 *The '99%-urp' Scheme and Evans' Second Counter-Example*

As I mentioned in chapter 1, Evans is not moved by potential rejoinders to his white-legged brown rabbit counter-example. Even if the '99%-urp' scheme manages to bypass Evans' first counter-example, Evans raises a further problem for the Quinean. Evans—see 1.7—considered the native sentence 'Parcial rojo gavagai' which according to the standard manual gets translated into English as 'There is a rabbit here and it is partly red'. We may nonetheless translate 'Parcial rojo gavagai'

à la Quine, Evans acknowledges, as 'There is a rabbit part here and it is wholly red'. Notice that 'There is a rabbit here and it is partly red' and 'There is a rabbit part here and it is wholly red' are stimulus synonymous. So far, so good. However, Evans goes on to argue that the Quinean proposal will loose its empirical adequacy as soon as we pay attention to more complex Native sentences. Consider the native sentence 'Parcial rojo, parcial verde gavagai', which we can standardly translate as 'There is a rabbit here and it is partly red and it is partly green'. The perverse rendering required, in order to preserve stimulus synonymy, Evans contends, would be something like 'There is a rabbit part here and it is wholly red and it is wholly green'. But such a translation clearly misrepresents Native usage, since it represents natives as assenting to 'Parcial rojo, parcial verde gavagai?' in presence of a rabbit part which is *both* wholly red and wholly green.

We saw in section 1.7 how a version of Hookway's disjunctive strategy could be successfully applied to cope with Evans' recalcitrant data. However, Hookway's strategy could only be applied at the expense of a greater complexity in the formulation of the axioms required to generate the appropriate satisfaction theorems. Fortunately for the Quinean we can once more make use of the '99%-urp' referential scheme to avoid Evans' second counter-example in a way which doesn't bring the structural complexity that Hookway's alternative implied. Consider the following extension of $PT_4$:

**$PT_4^+$**

Axioms:

**(e)** (x) (x satisfies 'gavagai' iff x is a 99%-urp)

**($e_3$)** (x) (x satisfies 'parcial'^$f$^Ø iff (∃y) (y is an undetached part of x & y satisfies $f$) & x satisfies Ø)

**($e_4$)** (x) (x satisfies 'rojo'^Ø iff (x is red & x satisfies Ø))

**($e_5$)** (x) (x satisfies 'verde'^Ø iff (x is green & x satisfies Ø))

**($e_6$)** (x) (x satisfies 'parcial'^*f*^'parcial'^*g*^Ø iff (∃y)(∃z) (y is an undetached part of x & z in an undetached part of x & x satisfies Ø & y satisfies *f* & z satisfies *g*))

Theorems:

**($e_7$)** (x) (x satisfies 'parcial'^'rojo'^'gavagai' iff (∃y) (y is an undetached part of x & y is red & x is a 99%-urp))

**($e_8$)** (x) (x satisfies 'parcial'^'verde'^'gavagai' iff (∃y) (y is an undetached part of x & y is green & x is a 99%-urp))

**($e_9$)** (x) (x satisfies 'parcial'^'rojo'^'parcial'^'verde'^'gavagai' iff (∃y)(∃z) (y is an undetached part of x & z is an undetached part of x & y is red & z is green & x is a 99%-urp))

Under $PT_4^+$ we translate the native sentence 'Parcial rojo gavagai' as in effect 'A 99%-urp here is partly red'. The reader can see that this perverse alternative preserves stimulus synonymy with respect to its standard counterpart 'A rabbit here is partly red'. Furthermore, when we move to more complex Native constructions such as 'Parcial rojo, parcial verde gavagai', we can see that $PT_4^+$ is not subject to the problems that Evans envisages. Following theorem ($e_9$), the appropriate perverse rendering of the native sentence is in effect 'A 99%-urp here is partly red and partly green'. Again, we can see that the perverse interpretation preserves stimulus synonymy with respect to its standard counterpart 'A rabbit here is partly red and partly green'. We would assent to/dissent from the perverse rendering in all those situations in which we would have assented to/dissented from the standard one. If it is legitimate of an object to be partly-*f* and partly not-*f*, the same goes for another

object composed of the 99% volume of the first. In short, we can perfectly be in presence of the partly red and partly green 99% undetached part of an object.[6]

We've seen how the perverse semantic theory $PT_4$ can cope with the counter-examples put forward by Evans. In the remainder of this chapter I shall elaborate on a different issue that will shed more light on the reasons why I believe that the '99%-urp' scheme is a good tool for the Quinean.

## 3.4 *Back to the Apparatus of Individuation*

Consider again the Native sentences 'Dos gavagai' and 'Dos rosas' which were translated standardly as 'There are (exactly) two rabbits' and 'There are (exactly) two roses' respectively—see chapter 1. By digging in the apparatus of individuation (plurals, identity, etc.)—see 1.4— the anti-Quinean hoped to discover some data recalcitrant to perverse interpretations of these sentences. Hookway managed to overcome potential difficulties by translating 'dos' in a context-dependent way. Following Hookway—see 1.4, ft. 18—we may equate 'dos' with 'two animals which are composed of', when dealing with rabbit-related utterances, and with 'two plants which are composed of', when dealing with rose-related ones. Once more, Hookway's strategy could be successfully applied, but at the cost of loosing structural simplicity when compared to ST.

---

[6] The reader should notice that the anti-Quinean cannot adduce in her favour a situation in which an object is 99% red and 1% green. The reason is that (e) has been defined in terms of *a* 99%-urp. Hence, if it were the case that *a particular* 99%-urp cannot be partly green because greeness applies to the 1% left of the whole rabbit, we are not in trouble: We've got plenty of other cases where the 1% of greeness applies to the surface of a 99%-urp.

Thanks to the '99%-urp' scheme, we do not need to make use of disjunctive rules of translation to deal with complex structures where the apparatus of individuation is present. Consider Quine's original rendering of 'gavagai' as contemplated under the perverse semantic theory $PT_1$ (see 1.6):

**(b)** (x) (x satisfies 'gavagai' iff x is an undetached rabbit part)

If we try to avoid Hookway's rendering of 'dos' (as related in each particular disjunct to animals or plants or minerals, etc.) and talk, instead, of satisfaction conditions over *things in general* by means of one non-disjunctive axiom, we are in trouble. The reason is that according to $PT_1$ we will obtain the following truth theorem for the native utterance 'Dos gavagai':

**(t)** 'dos'^'gavagai' is true iff $(\exists x)(\exists y)$ (x is an undetached rabbit part & y is an undetached rabbit part & $x \neq y$ & (z) (z is an undetached rabbit part $\rightarrow$ (z=x or z=y)))

Theorem (t) tells us that there are two, and no more than two, things which are undetached rabbit parts. But according to (t), native speakers would not assent to 'Dos gavagai?' even when faced with exactly two rabbits. For obviously, even a single rabbit has many more than two undetached parts.

However, we may substitute axiom (b) for (e)—i.e., the perverse axiom for 'gavagai' contemplated under $PT_4$—see 3.2:

**(e)** (x) (x satisfies 'gavagai' iff x is a 99%-urp)

By taking 'gavagai' as dividing its reference over 99%-urp, we obtain the following theorem:

**($t_1$)** 'dos'^'gavagai' is true iff $(\exists x)(\exists y)$ (x is a 99%-urp & y is a 99%-urp & $x \neq y$ & (z) (z is a 99%-urp $\rightarrow$ (z=x or z=y)))

This is better, but still won't do. Theorem ($t_1$) tells us that there are two, and no more than two, things which are 99%-urps. But, according to ($t_1$), 'Dos gavagai?' would still not be assented to in presence of a pair of rabbits: For each individual rabbit consists of indefinitely many 99%-urps, obtained by selecting a different 1% of the rabbit as the remainder.[7]

One final adjustment will permit us generate the truth theorem required. In order to preserve stimulus synonymy with respect to the standard semantic theory, we simply need the two 99% undetached rabbit parts not to overlap. Take the symbol '÷' to represent the fact that two objects are different in the sense that they share no particle at all. By changing 'y÷z' for 'y≠z', we shall obtain the following theorem:

**($t_2$)** 'dos'^'gavagai' is true iff $(\exists x)(\exists y)$ (x is a 99%-urp & y is a 99%-urp & x÷y & (z) (z is a 99%-urp $\rightarrow$ ($\neg$z÷x or $\neg$z÷y)))

[7] Notice that 'y≠z' in ($t_1$) just means that y is *different* from z. The disanalogy, however, could be simply a matter of not having one particle in common; y and z could, thus, be sharing the rest of their components.

Now, according to ($t_2$), we require exactly two rabbits, for we could not possibly be referring to two different 99% parts of one single rabbit which did not overlap.[8] Notice that thanks to the '99%-urp' semantic theory, we avoid deploying context-sensitive translations of 'dos'. The native term 'dos' can be translated as 'there are two non-overlapping ...'. Hence, we can couple the expression in question, 'dos', with the 99% undetached part of any object at all, irrespectively of its nature, avoiding, thus, having to discern among plants, animals, etc. In this way, according to the '99%-urp' scheme, we can translate the native sentence 'Dos rosas' (see 1.4) as 'There are exactly two non-overlapping 99% undetached rose parts'.

The reader might worry that $PT_4^+$ cannot assign 'Dos gavagai' the condition true if and only if there are exactly two rabbits. For, the reader might think, a single rabbit has indefinitely many (partially overlapping) 99%-urps. Hence the first rabbit provides an indefinitely large stock of 99%-urps none of which overlap with any of the indefinitey large number of 99%-urps provided by the second rabbit. However, $PT_4^+$ does get the truth conditions of 'Dos gavagai' right. Any choice of value for x and of a value for y rendering the sentence true selects a pair of non-overlapping 99%-urps, which perforce have to come one from each rabbit, and then there is no third non-overlapping 99%-urp. Thus 'Dos gavagai' comes out true if and only if there are two rabbits.

In conclusion, when we move from Evans' compounds of predicates to the apparatus of individuation, we can see that the '99%-urp' scheme still works. And furthermore, it avoids having to employ disjunctive rules of translation.

---

[8] Notice that it would have been useless to employ '÷' in (t) since y and z could share no particle at all, and still be two different things belonging to the same rabbit. We avoid this difficulty when the two things are as big as the 99% of a rabbit.

## 3.5 *Game-Theoretical Semantics*

We gain a further argument for the indiscernibility of $PT_4^+$ and ST, and for the superiority of $PT_4^+$ over Hookway's proposal, if we consider game-theoretical semantics as an epistemic model. ST, $PT_4^+$ and Hookway respectively provide the following translations of 'Dos gavagai':—[9]

(ST) $(\exists x)(\exists y)$ (x is a rabbit & y is a rabbit & $\neg x{=}y$ & $(z)$ ($\neg z$ is a rabbit $\vee$ ($z{=}x$ $\vee$ $z{=}y$)))

($PT_4^+$) $(\exists x)(\exists y)$ (x is a 99%-urp & y is a 99%-urp & $x \div y$ & $(z)$ ($\neg z$ is a 99%-urp $\vee$ ($\neg z \div x$ $\vee$ $\neg z \div y$)))

(H) $(\exists x)(\exists y)$ {Animal x & Animal y & $x \neq y$ & $(w)$ ($\neg w$ is a component of x $\vee$ w is an undetached rabbit part) & $(w)$ ($\neg w$ is a component of y $\vee$ w is an undetached rabbit part) & $(z)$ [$\neg$(Animal z & $(w)$ ($\neg w$ is a component of z $\vee$ w is an undetached rabbit part)) $\vee$ ($z{=}x$ $\vee$ $z{=}y$)]}

We usually think of a native assenting to 'Dos gavagai' in the obvious presence of a pair of rabbits, and hence the only relevant behaviour of the native might be immediate assent. But epistemic circumstances might be more difficult—the native might be set the task of finding out whether there are exactly two rabbits living in the large overgrown orchard, which might involve crawling around finding rabbits and distinguishing them from the other inhabitants of the orchard. We might then

---

[9] I have replaced expressions of the form '$p \rightarrow q$' by '$\neg p \vee q$', since, as we shall shortly see, Hintikka does not give game-rules for '$\rightarrow$'.

expect more complex behaviour leading up to an assent to 'Dos gavagai?', behaviour which displays a canonical verification procedure following the logical form of (ST), or ($PT_4^+$), or (H). Hintikka's game-theoretical semantics gives us a model for canonical verification of 'Dos gavagai' under our three proposed translations. In this way we might look for behavioural evidence in favour of one or other translation.

In *Logic, Language-games and Information*, Hintikka offers game-theoretical semantics which we can apply to (ST), ($PT_4^+$), and (H)[10]—see Hintikka, 1973, pp. 86-8. Simplified, the game of 'searching and finding' goes as follows:

The game is played on a given quantified sentence, S. The game is played by two persons—the truth proponent of S (hereafer the proponent) who is committed to showing that S is true, and her opponent, the falsity proponent of S (hereafter the opponent), who is committed to showing that S is not true. Proponent and opponent are invited to play out semantic games on S, according to the rules set out below. At each round of a game the play focuses on the main constant and results in a simpler sentence, which is then the subject of play in the next round of the game, until an atomic formula is reached when the game stops. If the atomic formula is true, whoever is proponent at that stage of the game has won, and if it is false whoever is opponent at that stage of the game has won. For S to be true is for the proponent of S to have a winning strategy. That is, a repertoire of plays such that she wins whatever her opponent may play. The interesting idea from our point of view is that a winning strategy will reflect the logical form of S, since plays of the game will

[10] The reader may care to consult Hintikka (1976) for an employment of game-theoretical semantics in a context wider than radical translation as a way to grasp the connection between quantifiers of Formal Logic and quantifiers in Natural Languages. See also Tennant (1987).

trace the nested structure of logical constants in S. Hence, we will expect, the behaviour of one who is seeking to discover whether she has a winning strategy on S will, in general, reflect the nested logical structure of S, since the players have to discover whether or not they have winning strategies on various simpler sentences generated in the play on S when the logical constants are succesively stripped away. Thus we may hope to predict behavioural differences in between one who is a proponent of (ST) as against one who is a proponent of ($PT_4^+$), as against one who is a proponent of (H). At least we may hope to do so when the determination of 'Dos gavagai' is particularly difficult and forced to follow an ideal canonical epistemic route mapped out by its logical form.[11, 12]

To play the game we need to learn some basic rules. At each stage of the game, at which a quantifier is the main constant, a player chooses a member of the universe of discourse. Similarily, at each stage at which a propositional operator is the main constant, a player chooses a disjunct or a conjunct, depending on the form

---

[11] The relevancy of Hintikka's strategy for our purposes is that the games are played in strict behavioural terms, without appeal to normative or rational considerations—see 1.2. Although Hintikka's concern is not the translation of terms and ontologies, but rather the translation of quantifiers—see Hintikka (1973), pp. 87-ff.—I believe, nonetheless, that we can employ his insights to throw some light upon our current semantic and ontological worries.

[12] By 'ideal' I mean the following: In any particular game, the number of rounds necessary to arrive at an atomic sentence and verify it depends on the *ability* of both contestants. The fact that a given sentence is true doesn't imply that it *will* be verified by the proponent, but only that it *can* be verified. Whether the proponent manages to verify it or not depends on how smart she is in her choice of individuals. In the same way, if her opponent is dumber, it will be easier for the proponent to win; but if the opponent plays a good game, the proponent will have to perform at her best to win the game. Hence, what I mean by an 'ideal game' is that game where the two contenders play at the possibly *maximum* level to achieve their purposes.

of the sentence being considered. But we need to know *who* is the one to choose. This will depend on the kind of sentence in question. Hintikka gives the following five rules:

**$R_1$** If S is $(\exists x)$ F(x), the proponent chooses a member of D—i.e., the universe of discourse—, and gives it a proper name, say 'b'. The game is then continued with respect to F(b).

**$R_2$** If S is (x) F(x), the same happens except that the opponent chooses b.

**$R_3$** If S is $(F \vee G)$, the proponent chooses F or G, and the game is continued with respect to it.

**$R_4$** If S is $(F \wedge G)$, the same happens except that the opponent makes the choice.

**$R_5$** If S is $\neg F$, the roles of the two players (as defined by rules $R_1$, $R_2$, $R_3$ and $R_4$) are reversed and the game is continued with respect to F. (Adapted from Hintikka (1976), p. 217)

By following these rules, the proponent and her opponent will keep on choosing individuals, disjuncts and conjuncts alternatively (depending on the form of the sentence S) until they obtain an atomic sentence which contains no quantifier phrase at all. If that atomic sentence is true then whoever has the role of proponent at that stage wins, and otherwise whoever has the role of opponent at that stage wins. Now we can see why Hintikka calls it a game of 'seeking and finding'. Each player seeks for the individuals that will verify or falsify any quantified statement in dispute, or seeks which disjunct or conjunct to select. The underlying thought in Hintikka's strategy is then that, for decidable statements, if S is true, then the proponent of S will have a winning strategy to verify it.

Let's now see the bearing of Game-Theoretical Semantics for our present purposes. Recall that our three rival translation manuals offer the following as the logical form of 'Dos gavagai':—

(ST) $(\exists x)(\exists y)$ (x is a rabbit & y is a rabbit & $\neg x{=}y$ & (z) ($\neg$z is a rabbit $\vee$ (z=x $\vee$ z=y)))

($PT_4^+$) $(\exists x)(\exists y)$ (x is a 99%-urp & y is a 99%-urp & $x \div y$ & (z) ($\neg$z is a 99%-urp $\vee$ ($\neg z \div x \vee \neg z \div y$)))

(H) $(\exists x)(\exists y)$ {Animal x & Animal y & $x \neq y$ & (w) ($\neg$w is a component of x $\vee$ w is an undetached rabbit part) & (w) ($\neg$w is a component of y $\vee$ w is an undetached rabbit part) & (z) [$\neg$(Animal z & (w) ($\neg$w is a component of z $\vee$ w is an undetached rabbit part)) $\vee$ (z=x $\vee$ z=y)]}

As a preliminary and to fix ideas, I illustrate by describing a game on (ST), with obvious abbreviations.

**(s)** $(\exists x)(\exists y)$ ((Rx & Ry) & $x \neq y$ & (z) ($\neg$Rz $\vee$ (z=x $\vee$ z=y)))

Round 1:[13] Sp chooses $r_1$,

Play continues on:—

$(\exists y)$ (($Rr_1$ & Ry) & $r_1 \neq y$ & (z) ($\neg$Rz $\vee$ (z=$r_1$ $\vee$ z=y)))

[13] 'Sp' stands for the proponent of (s), and 'So' for her opponent. For economy I take the set of individuals on which the predicates are interpreted to contain only three objects: two rabbits and an unspecified object other than a rabbit—abbreviated respectively $r_1$, $r_2$, and o.

Round 2: Sp chooses $r_2$,

Play continues on:—

$$(Rr_1 \,\&\, Rr_2) \,\&\, r_1 \neq r_2 \,\&\, (z)\,(\neg Rz \vee (z=r_1 \vee z=r_2))$$

Round 3: So chooses 3rd. conjunct,

Play continues on:—

$$(z)\,(\neg Rz \vee (z=r_1 \vee z=r_2))$$

Round 4: So chooses o,

Play continues on:—

$$\neg Ro \vee (o=r_1 \vee o=r_2)$$

Round 5: Sp chooses 1st. disjunct,

Play continues on:—

$$\neg Ro$$

Round 6: So is committed to the truth and Sp to the falsity of:—

$$Ro$$

Game Over

Sp wins if 'Ro' is false, otherwise So has won this particular game.

If (ST) gives the logical form of 'Dos gavagai' then one who asserts 'Dos gavagai' claims, in effect, to have a winning strategy on (ST). So we may expect the behaviour leading up to an assertion of 'Dos gavagai' to be, in an ideal case, the behaviour of one seeking to discover whether they have a winning strategy on (ST). And similarily, of course, for ($PT_4^+$) and (H). We may now note a striking parallel between (ST) and ($PT_4^+$).

For every game on (ST) leading to an atomic sentence in the left hand column, there is an exactly parallel game on ($PT_4^+$) leading to the 'atomic' sentences in the right hand column:—

| | |
|---|---|
| A is a rabbit | A* is a 99%-urp |
| B is a rabbit | B* is a 99%-urp |
| A=B | A÷B |
| C is a rabbit | C* is a 99%-urp |
| C=A | ¬C*÷A |
| C=B | ¬C*÷B |

where A=A*, unless A is a rabbit in which case A* is a 99% undetached part of that rabbit, and B and B*, and C and C*, similarly.

Now, if the game on (ST) produces a win for the proponent, so does the corresponding game on ($PT_4^+$), and *vice versa*. So it seems that the behaviour of a proponent trying to see whether they have a winning strategy on (ST) will be indistinguishable from the behaviour of a proponent trying to see whether they have a winning strategy on ($PT_4^+$).

However, it might seem that nonehteless there are two differences, which I will consider in turn:—

(1) In the last two cases, games on ($PT_4^+$) have a further round in which roles are swapped and a final round is played on 'C*÷A' or on 'C*÷B'. Perhaps we can hope to test this difference of length in their respective games behaviouristically. But this is not a difference which registers in behaviour. The proponent is the asserter of 'Dos gavagai', but the opponent is only a notional character. All that happens is that when a proponent reaches 'C=A' she has to determine whether C and A are identical. Likewise, all that happens when a proponent reaches '¬C*÷A' is that they have to determine whether C* and A partially overlap. No behaviour will reveal which of these tasks a proponent is engaged in. Similarly for 'C=B' and '¬C*÷B'.

(2) 'A is a rabbit' is an atomic sentence, and it is assumed that when a game is played in which this is the terminus, and proponent and opponent know who has won, this is because 'A is a rabbit' is verified or falsified by direct observation. But 'A* is a 99%-urp' is not, in absolute terms, an atomic sentence. It has significant semantically relevant structure. Thus, it is to be distinguished from, for example, 'A* is a 5%-urp'. So it might seemt that we should analyse 'A is a 99%-urp' along the lines of

$$(\exists x)(\exists y)(\exists n)\ (x \text{ is a rabbit } \&\ y=A^* \ \&\ n=99 \ \&\ y \text{ is } n\% \text{ of } x),$$

and then the game should continue on this. However, this is to misunderstand the nature of Quine's proposed indeterminacy of radical translation, and the proposal

($PT_4^+$) in particular. Although 'A* is an undetached rabbit part' is indeed semantically complex, Quine assumes that it is *epistemically* equivalent to 'A is a rabbit'. On all occasions in which one is able to verify or falsify 'A is a rabbit' by direct observation, one can also verify or falsify 'A* is an undetached rabbit part' by direct observation, and *vice versa*, Quine assumes. The same holds for 'A is a rabbit' and 'A* is a 99%-urp', we are assuming. So from the point of view of epistemic behaviour, we can regard games which reach 'A* is a 99%-urp' as terminating there, as we do regard games which reach 'A is a rabbit', the winner being decided by direct observation.

Thus, in sum, any behaviour which is interpretable as seeking and finding in the service of discovering a winning strategy on (ST) is equally interpretable as seeking and finding in the service of discovering a winning strategy on ($PT_4^+$), and *vice versa.*

Unfortunately for Hookway's route, the same cannot be said for (ST) and (H). Recall the logical form of 'Dos gavagai' offered by Hookway's translation manual:

(H) $(\exists x)(\exists y)$ {Animal x & Animal y & $\neg x=y$ & (w) ($\neg$w is a component of x $\vee$ w is an undetached rabbit part) & (w) ($\neg$w is a component of y $\vee$ w is an undetached rabbit part) & (z) [$\neg$(Animal z & (w) ($\neg$w is a component of z $\vee$ w is an undetached rabbit part)) $\vee$ (z=x $\vee$ z=y)]}

As we saw above, games on (ST) lead to one or other of:—

A is a rabbit
B is a rabbit
A=B

C is a rabbit

C=A

C=B

On the other hand, games on (H) lead to one or other of:—

A is an animal

B is an animal

A=B

C is a component of A

C is an undetached rabbit part

D is a component of B

D is an undetached rabbit part

E is an animal

F is a component of E

F is an undetached rabbit part

F=A

F=B

A sympathizer of Hookway who asserts 'Dos gavagai' would claim, in effect, to have a winning strategy on (H)—assuming that (H) gives the logical form of 'Dos gavagai'. As in the cases of (ST) and ($PT_4^+$) above, we may expect the behaviour leading up to an assertion of 'Dos gavagai' to be, in an ideal case, the behaviour of one seeking to discover whether they have a winning strategy on (H). However, the reader can see that unlike games on ($PT_4^+$), games on (H) lead to one or other of the above sentences by routes which are *not* images of those on (ST). This disanalogy permits us to predict behavioural differences in between one who is a proponent of (ST), as against one who is a proponent of (H). We shall be able to distinguish the behaviour of a proponent trying to see whether they have a winning

strategy on (ST) from the behaviour of a proponent trying to see whether they have a winning strategy on (H). Therefore, any behaviour which is interpretable as seeking and finding in the service of discovering a winning strategy on (ST)—or for that matter, on ($PT_4^+$)—*cannot* be interpreted as seeking and finding in the service of discovering a winning strategy on Hookway's route, (H).

Although we have only considered one example, 'Dos gavagai', the points made generalize. There is an obvious isomorphism between the translation manuals (ST) and ($PT_4^+$) with 'is a rabbit' in (ST) as the image of 'is a 99%-urp' in ($PT_4^+$). Likewise, there is an obvious lack of isomorphism between the translation manuals (ST) and ($PT_4^+$), on the one hand, and (H), on the other. Provided we can take 'is a rabbit' as observationally equivalent to 'is a 99%-urp'—see 3.2, above—, then the native's behaviour when seeking to verify a native sentence S will be equally interpretable as seeking to verify that she has a winning strategy on sentence S delivered by (ST), and as seeking to verify that she has a winning strategy on the corresponding sentence delivered by ($PT_4^+$), and *vice versa.*

In sum, by looking at the native's complex patterns of behaviour leading up to an assent to 'Dos gavagai?', I contended, we've gained a further argument for the indiscernibility of the semantic theories $PT_4^+$ and ST, and for the superiority of $PT_4^+$ over Hookway's proposal. And plausibly, the points made concerning 'Dos gavagai' generalize to all sentences of Native—see chapters 4 and 5, below.

## 3.6 *Conclusion*

In this chapter we've seen how the Quinean can be semantically perverse with no need to make baroque adjustments in terms of the derivational structure of her

perverse theory. This renders the '99%-urp' scheme of reference a promising candidate to exemplify the Inscrutability Thesis; especially taking into account that research in the neurosciences might end up backing Evans' argument against structurally complex semantic theories—although see chapter 4, below.

Unfortunately for the Quinean, structural simplicity is not the only front that endangers perverse semantic theorizing. Crispin Wright (1997) has recently argued that apart from structural simplicity, the Quinean faces bigger worries. Wright is thinking of simplicity, not in the derivational structure of the perverse theories, but in the *psychological* theory that accompanies them. Psychological simplicity, as we shall see next, can become a powerful weapon for the anti-Quinean to exploit.

# 4

# *A CONNECTIONIST DEFENCE OF THE INSCRUTABILITY THESIS*

## 4.1 *Introduction*

In chapters 1 and 3 we saw how two different perverse semantic theories—$PT_2$ and $PT_4$—could be developed in order to preserve their empirical adequacy against our 'privileged' standard semantic theory, ST. All the hurdles, though, for these perverse alternatives consisted of behavioural and hypothetical neurophysiological data, and considerations regarding *structural* simplicity may tip the balance against perverse interpretations of Native. In this chapter I take up a new challenge for the defender of the Inscrutability Thesis. The threat comes this time, not from considerations regarding complexity in the derivational structure of the Quinean perverse candidates, but from the complexity in the *psychological* theory that

accompanies semantic theorizing in general. The challenge is, in my opinion, far more serious than those tackled in previous chapters. In order to address it I shall elaborate on current issues in Connectionist Theory, producing then, I hope, a *Connectionist Defence of the Inscrutability Thesis.*

Before getting started, let me briefly outline the programme of this chapter. In section 4.2, I shall introduce and develop a criterion recently produced by Wright (1997) in terms of 'psychological simplicity' which threatens the perverse semantic proposal I offered in chapter 3. In section 4.3, I shall argue that a Language-of-Thought—LOT—model of human cognition could motivate Wright's criterion, favouring thus a standard interpretation of Native along the lines reviewed in chapter 1. In sections 4.4-4.6 I shall introduce the reader to some basic aspects of connectionist theory, and elaborate on a particularly promising neurocomputational approach to language processing put forward by Jeff Elman (1992; 1998). I shall argue that if instead of endorsing a LOT hypothesis, we model human cognition by a *recurrent* neural network à la Elman, then Wright's criterion is unmotivated. In particular, I shall argue that considerations regarding 'psychological simplicity' are *neutral*, favouring neither a standard interpretation of Native, nor a perverse one. In section 4.7, I shall consider two lines of response to my connectionist defence of the Inscrutability Thesis. I shall rejoin to one of them, deferring full treatment of the other line of response until chapter 7, where I'll look in more detail to some recent neurocomputational research in order to answer it. In section 4.8 I shall argue that connectionism can account for the systematicity and compositionality of thought while avoiding a symbolic—LOT—implementation. This is an important result that

will permit us address one of the two caveats left unanswered in chapter 2 (section 2.3). Finally, I shall close the discussion in 4.9 by addressing a minor worry raised by a sympathizer of connectionism. In addition, I shall give some hints as to how connectionism may fit with the thesis of Semantic Holism—another pivotal thesis of Quine.

## 4.2 *Wright's 'Psychological-Simplicity' Argument*

Crispin Wright (1997) has recently proposed a line of argument against the Inscrutability Thesis which focuses upon the *conceptual repertoire* of native speakers. Wright's overall argument does not rely on the considerations regarding 'structural simplicity' that I addressed in chapter 2. Instead, Wright contrasts the simplicity of the conceptual repertoire imputed to the native by the standard manual with the contrasting complexity of the conceptual repertoire imputed to the native by a perverse manual. His aim is to make use of some sort of 'psychological-simplicity' criterion in order to discredit any perverse semantic theorizing. Wright targets Hookway's perverse semantic proposal, which employs disjunctive axioms—see 1.6—, but his argument, if valid, applies equally to $PT_4$—the perverse semantic proposal I advanced in chapter 3 (see 3.2). The reason is that even though a perverse semantic theory which for example divides the reference of 'gavagai' over 99%-urp, rather than over rabbits, is as simple, structurally speaking, as the standard one is—see 3.2—, it is nonetheless true, or so Wright believes, that such a theory imputes a great deal of *psychological complexity* to the native (Wright, 1997,

p. 412). And now, Wright contends, if rival semantic theories impute different conceptual repertoires to natives, but one imputes a simpler repertoire than the others, then that one is *objectively* speaking the correct semantic theory. Hence, the standard theory ST—see 1.6—is the only correct semantic theory.[1]

Let us elaborate on Wright's argument to see if it poses a serious threat to the Quinean. Wright claims that

> **(A)** the basic clauses of our semantic theory are to assign reference and satisfaction-conditions in ways which are *presumed to correspond to the conceptual repertoire of speakers of the language in question.*[2] (*Ibid.*, p. 412, Wright's emphasis)

It may seem that Wright begs the question against Quine. Obviously, a hard-line

---

[1] Someone may argue that the 'psychological complexity' imputed to the native by a perverse semantic theory will eventually show up in complex patterns of behavioural dispositions. In this way, loss or acquisition of, say, linguistic dispositions under certain circumstances are *observable* higher-order dispositions which may act as a constraint, tipping the balance against perverse interpretations of Native. (This line of argument was prompted by an anonymous referee of *Mind and Language* in response to Calvo Garzón, 2000a). This, however, should not cause any concern for, as the discussion in chapters 2 and 3 revealed, the '99%-urp' referential scheme would conform to such a constraint. Wright's attack, thus, must come from a different corner, as we'll see next.

[2] The reader should notice that by agreeing on this point we're not being committed to accepting Evans' *Mirror Constraint* (see 2.2). Wright's contention has nothing to do with *mirroring* the derivational structure of our semantic theories—see 2.4, fn. 11. Rather, according to Wright's psychological approach, speakers' conceptual repertoires must mirror our semantic theories' *basic clauses*, not the routes, departing from them, which generate the various semantic theorems.

Quinean would not accept Wright's premise, since it trades in *concepts*. However, we can interpret the premise in a way acceptable to a Quinean. The idea is to naturalize concepts in such a way that Wright's 'conceptual repertoire' can be transposed into a form which the Quinean should admit as legitimate—see 3.3. Whatever naturalizing strategy we adopt—see 5.3-5.6 below—the key point, scientifically speaking, is that we will require some relation between the concepts belonging to a speaker's conceptual repertoire, expressed by words, and the information content of real internal states in the brain. So, assuming there is such a relation, Wright's premise should be accepted by a Quinean. Wright's argument can then proceed as follows. Firstly, Wright notices, with respect to putative perverse alternatives to the standard scheme, that

> **(B)** the range of concepts necessary in order to formulate their various [basic] clauses in each case includes, but is not included in, the simple range of concepts of observable spatio-temporal continuants and their observable properties which the favoured scheme deploys. (*Ibid.* p. 412)

Taking for example the Quinean perverse schemes that divide the reference of 'gavagai' over undetached parts of rabbits or over their temporal stages respectively, Wright argues:

> **(C)** To have the concept of an undetached rabbit part, you need a concept of the integrated individual of which such parts are parts; to have the concept of a temporal stage of a rabbit, you need to grasp the idea of the spatio-temporal continuant of which such a stage is a stage. (*Ibid.* p. 412)

If we add as a *manifestation requirement* that the basic clauses should not assign to a speaker the possession of a larger repertoire of concepts than is needed to explain the subject's behaviour, we can see that Wright's argument poses a threat to the perverse semantic theory I offered in chapter 3.[3]

However, we need to guard against a misreading of the above argument whose clarification will prove crucial for my purposes in due course (see chapter 5 below). We are considering the conceptual repertoires assigned by the *basic* clauses of the standard and the perverse semantic theories respectively. The *total* conceptual repertoire of the native speakers will of course include all the complex concepts

---

[3] To keep the record straight, it must be noted that Wright reinforces his argument not with the aforementioned 'manifestation requirement', but rather with the following methodological caveat: "that the conceptual repertoire which radical interpretation may permissibly ascribe to speakers should exceed what is actually expressible in their language, as so interpreted, only if its ascription to them is necessary in other ways in order to account for their linguistic competence" (Wright, 1997, p. 412). Wright (personal communication) acknowledges that it is unclear how the methodological constraint he offers, as it stands, would deliver him the goods. Notice that unless some further psychological or neurophysiological explanation is forthcoming as to *why* speakers cannot be ascribed a conceptual repertoire which is not strictly necessary to explain their linguistic competencies, Wright's methodological constraint collapses into the methodological considerations we reviewed in chapter 2, and is thus doomed for the reasons I offered there. I am happy to accept that a developed version of Wright's constraint, or of the manifestation requirement sketched above, may well play the role Wright desires—indeed, Wright's (1992) 'Wide Cosmological Role' constraint may well be a good candidate. However, I would need to see a proposal along those lines in some detail before submitting it to critical scrutiny. Nevertheless, we need not worry about this

which they can build from the simple lexicon of Native by the usual combinatorial means. In general, *total* sets of concepts will be the same under the perverse and the standard theories. We can see this by transferring Quine's case of Radical Translation to Home.[4] Suppose we are devising translation manuals for fellow speakers. I may translate your English sentence 'There is a rabbit' homophonically as my 'There is a rabbit'. Or I could translate it heterophonically as my 'There is a 99%-urp'. Since my sentence 'There is a 99%-urp' is a well-formed sentence of English, it is one you could produce and, hence, must be subject to translation into my English. Again, my homophonic manual would equate it with my 'There is a 99%-urp', whereas my heterophonic manual would translate it as 'There is a 99% undetached part of a 99%-urp'. Once again, this sentence is also a well-formed sentence in your English. So, once again, I need to translate it and can do so either via my standard manual or via my perverse manual. Obviously the process iterates indefinitely. The point of all this is that the total conceptual repertoire assigned via either manual is the same. Hence, Wright's argument should be taken to concern only the conceptual repertoire imputed by the *basic* clauses of the rival translation

---

issue for our present interests, since my counter-arguments in this chapter hinge somewhere else, calling into question the core of Wright's argument—i.e., quotes (B) and (C) above.

[4] Setting the parable of Radical Translation in a *home* environment—i.e., English-to-English translation— should not alter matters significantly, and Wright would agree. The success of the Inscrutability Thesis cannot be dependent on the object-language being *inferior*—grammatically and/or semantically speaking—with respect to the home language. Otherwise, the Inscrutability of Reference would amount to no more than a trivial—as far as Semantics is concerned—clash of cultures (see chapter 1).

manuals.

Wright reads the basic clauses realistically—as we saw in the quote labeled (A) above. Hence, he takes the conceptual repertoire of the basic clauses to be subject to a manifestation requirement—although see fn. 3 above. A Quinean may seek to naturalize the facts recorded by the basic clauses in either of two ways: As a LOT hypothesis or in a Connectionist architecture. We may then ask for manifestable evidence in favour of one or the other semantic theory. The question of which semantic theory is correct becomes subsumed, I claim, under the question of which account of the brain's architecture is correct. I shall argue below that a LOT hypothesis favours ST over $PT_4$, whereas a connectionist setting is neutral between ST and $PT_4$. The remainder of this chapter will be devoted to developing this argument.

## 4.3 *Psychological Simplicity and The Syntactic Image*

What structure do the *representations* in the brain have?[5] In other words, how is

[5] For strategical reasons—see chapter 7, below—I assume throughout the remainder of my dissertation a *representationalist* approach to cognition both in the classical and the connectionist theoretical frameworks. Although this may sound somewhat platitudinous from within the classical approach, in the second case it is less than obvious. Those keen on eliminating content *altogether* may care to consult, for example, Beer (1995), Freeman and Skarda (1990), Keijzer (1998), Ramsey (1993), and van Gelder (1995) for illustrations of how connectionist networks can perform particular tasks with no need for viewing the *internal* apparatus as representational. For some key cases that

information encoded in a cognitive system? In Fodor's view, the study of higher cognitive abilities—thought, language mastery, etc.—and, in particular, explaining the *systematicity*, *productivity* and *inferential coherence* of thought processes, requires a *symbolic* treatment.[6] And the *best* metaphor at hand for the way information is encoded is human language. What we then have is a *linguiform* structured cognitive system. The underlying idea is that thinking can be seen as logic-like inferential processing—i.e., some sort of sentence-crunching. In Fodor and Pylyshyn's (1988) view, systematicity, productivity and inferential coherence can only be explained from a LOT perspective—see below. Let us elaborate.

According to a LOT hypothesis we, as thinkers, have the capacity to entertain thoughts with particular contents which are carried by the mental representations of LOT. For example, to entertain the thought THERE IS A WHITE 99%-URP OVER THERE,[7] is for us to be related to a *mental representation* carrying that particular content. In Fodor's view, concepts are word-types of LOT, and our employment of

---

pose a problem to the anti-representationalist—Andy Clark dubs these cases 'representation-hungry problems'—see Clark, 1997, chapter 8, and Clark and Toribio (1994).

[6] Put bluntly, we say that thought processes are systematic to the extent that our capacity to entertain or grasp the thought *AB* is directly connected with our capacity to entertain or grasp the thought *BA*. Thoughts, furthermore, are productive in the sense that we have the ability to entertain or grasp an *indefinite* number of increasingly complex thoughts: *A*, *AB*, *ABC*, ...—although neurophysiological constraints on human *hardware* capacities will unavoidably kick in. And lastly, human thought is inferentially coherent since our entertaining, or at least our grasping, the thought *A&B* triggers our grasping the thought *A* and the thought *B*.

[7] From now on I shall use capital letters to express concepts.

concepts is the occurrence of word-tokens of LOT.[8] In this picture, *context-independence* is a key feature. Fodor (1987, p. 137) notes that the constituent 'P' in the formula 'P' is a token of the same representational type as the 'P' in the formula 'P&Q', if 'P' is to be a consequence of 'P&Q'. Mental representations are formed out of context-independent constituents in such a way that constituents appear in different thoughts as syntactically identical tokens with the same conceptual content. I shall refer to this kind of context-independence, as *Classical Constituency.* In short, LOT and its classical form of constituency amount to claiming that:

> (1) (some) mental formulas have mental formulas as parts; and
> (2) the parts are 'transportable': the same parts can appear in *lots* of mental formulas. (*Ibid.*, p. 137)

Classical constituency, I contend, motivates Wright's 'psychological simplicity' argument. The working hypothesis of LOT is that there must be some causal relation between the speakers' strings of LOT and the strings of English which reflects a syntactic similitude between LOT and English strings. Fodor and

---

[8] See Fodor (1975; 1987). There are different versions of the LOT hypothesis—the reader may care to consult for example Field (1978), and Harman (1973). For the earliest explicit treatment of the LOT hypothesis, see Sellars (1968). Some people maintain that LOT is *actually* the thinkers' spoken language, but internalized. Others take LOT to be the analog of a hidden machine code. We do not need to decide which is the most plausible. We just need to pay attention to a key feature of LOT models: Classical Constituency (see below). For a quick appraisal of some problems that the *LOTTER* faces see Clark (1994).

Pylyshyn understand quite literally the linguiform metaphor of thought-processes:

> [The] symbol structures in a Classical model are assumed to correspond to real physical structures in the brain and the *combinatorial structure* of a representation is supposed to have a counterpart in structural relations among physical properties of the brain. (Fodor and Pylyshyn, 1988, p.13)

And, Fodor and Pylyshyn continue,

> the relation 'part of', which holds between a relatively simple symbol and a more complex one, is assumed to correspond to some physical relation among brain states. (*Ibid.*, p.13)

In this way, if the perverse scheme assigns to 'gavagai' the *phrasal* concept 99%-URP—expressed by a lexically complex phrase of English ('99%-urp')—, and we apply the linguiform analogy quite literally, we can see why this phrasal concept contains, among others, the atomic concept RABBIT. Because in the corresponding strings of LOT, the token RABBIT of LOT occurs in any token of 99%-URP of LOT. Therefore, we can see why Wright's argument holds. Employing the phrasal concept 99%URP involves employing some *word-tokens of LOT of the same word-type*—i.e., RABBIT. In short, we shall not be able to entertain for example a 99%-URP-related thought without *exercising* the concept of a rabbit, among others. The lexical concept RABBIT is, thus, *psychologically simpler* than the phrasal concept 99%-URP.

By approaching the issue of the naturalization of concepts from a LOT perspective, we've seen how Wright may hold to his principle of psychological

simplicity, and hope to put the Quinean up against the ropes. The story, however, as we are about to see, looks rather different once we approach the issue from a non-classical, connectionist, perspective where constituency is *non-classical* in a sense yet to be explained.[9]

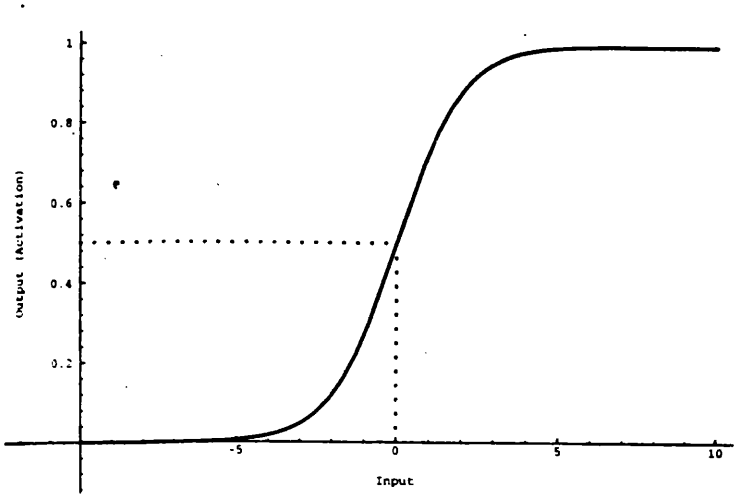
## 4.4 *Basic Aspects of Connectionism: Components and Dynamics*

In this section I shall review the basic components and dynamics of connectionist networks. Connectionism offers a new 'biologically- and developmentally-inspired approach' to cognition—see Elman *et al.*, 1996, chapter 2—which differs critically from the *Syntactic Image* championed by classical cognitive scientists—see 4.3 above, and the references given there. Computations, in the connectionist guise, are based primarily on the interconnection of many simple units whose dynamics seek to explain complex patterns of behaviour, whilst avoiding recurring to the explicit symbols and algorithms that classical computationalism relies on. The basic components of a connectionist network are simple processing units and connections between those units. Units—the reader may think of them as '*toy*-neurons'—have either a binary level of activation (0 or 1), or a range of values varying between 0

---

[9] The arguments I shall elaborate on in the remainder of this chapter, and in chapters 5 and 7, rely pretty heavily upon some key features of connectionist theory. I shall thus spend some time in sections 4.4 and 4.5 to introduce the reader to some basic aspects of connectionism. For philosophically-oriented introductions to connectionist theory, the reader may care to consult Bechtel and Abrahamsen (1991), Clark (1989; 1993), or Tienson (1988). Those familiar with the basic tenets of connectionism may wish to skip sections 4.4 and 4.5, and jump ahead to section 4.6.

and 1. Units receive input signals from other units—or from the environment, in the case of input units—via connections of various weights and polarities. The weights take the form of real-valued numbers, and indicate the strengths of the connections among the units. To obtain the input value to a unit *i* from a sending unit *j*, we multiply the activation value of unit *j* by the weight of the connection between unit *j* and unit *i*. Then the activation values from all units inputting to unit *i* are summed determining thus the *netinput* to unit *i*. In this way, units can be either excited or inhibited as a function of the existing connections and their values, and as a result, they acquire new levels of activation which may result in the emission of an output signal of a certain strength. The output signal emitted by a unit need not coincide in value with the netinput to that unit. Rather, output activation values are the result of an arithmetical function performed on the netinput.[10] The networks I shall make use




[10] Units whose output activations do coincide with their netinput are called *linear* units. However, for our purposes (see 4.8 below) we're interested in *non-linear* activation functions.

[Fig. 4.1]: The sigmoid activation function often used for units in neural networks. Outputs (along the ordinate) are shown for a range of possible inputs (abscissa). Units with this sort of activation function exhibit an all or nothing response given very positive or very negative inputs; but they are very sensitive to small differences within a narrow range around 0. With an absence of input, the nodes output 0.5, which is in the middle of their response range.[11] (from Elman *et al.*, 1996, p. 53)

of deploy the following logistic (sigmoid) activation function:[12]

$$a_i = 1/1+e^{-net_i}$$

where $a_i$ stands for the output value of unit *i*; $net_i$ for the net input to unit *i;* and e is the exponential. The graph above shows the output value $a_i$ of unit *i* for any given netinput.

We can now see how units *connect* to each other to form the skeleton of the network. In a *simple feedforward network*[13]—see fig. 4.2—units are organized into

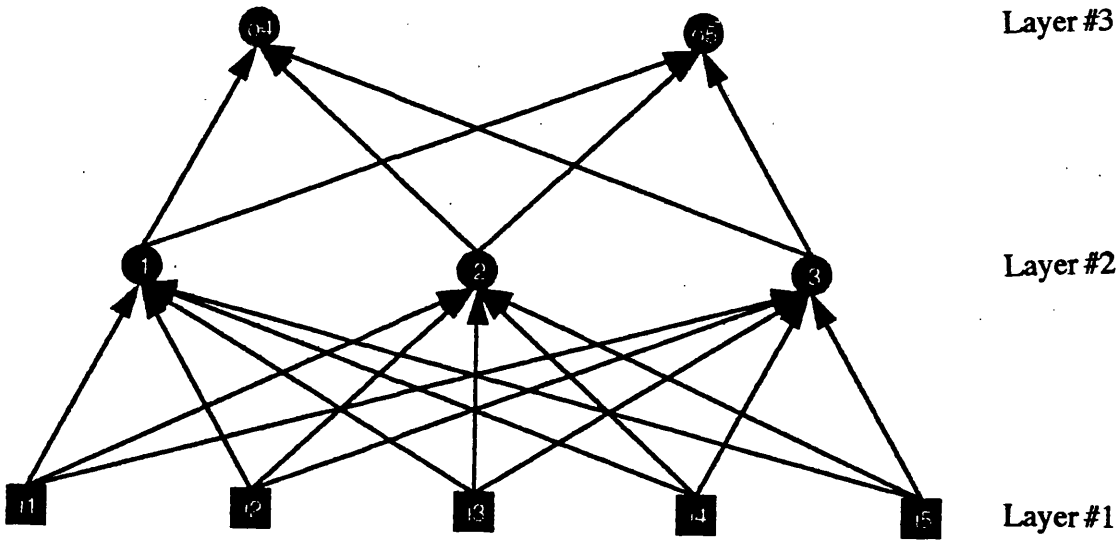
---

[11] This sort of non-linear response will be crucial for our purposes since, to advance a key point, it allows networks, which are not governed by explicit rules, to behave in a rule-like manner—see 4.8 below.

[12] The logistic activation function is an activation rule which takes into account parameters such as the netinput, the previous activation value, the decay rate, etc. We need not be bothered with the details here. The reader interested in the fine-grained mechanisms is urged to visit the *locus classici* (the 'bible' of connectionist theory): —Rumelhart, McClelland et al. (1986), and McClelland, Rumelhart et al. (1986). Bechtel and Abrahamsen (1991), and Elman *et al.* (1996) offer as well exhaustive introductions to the mechanics of neural networks.

[13] There are many different sorts of connectionist architecture to be found in the recent literature, though we can ignore for our purposes all such diversity—see Bechtel and Abrahamsen (1991) for

at least three separate layers. One input layer, one or more hidden ones, and one output layer. The units of a given layer connect exclusively to units of the next layer. In this way, activations feed forward from the input to the output layer passing through the hidden layer/s. The activation pattern of any of the layers corresponds to the sequence of activation values of the units that form the layer. We can treat these patterns of activation as vectors in a n-dimensional space, where *n* corresponds to the number of units that constitute the layer.




[Fig. 4.2]: A simple feedforward network. Units in layer 1 are input units. Units in layer 3 are

---

an exposition of some of the most relevant types. See also Hanson and Burr (1990) for a sketched taxonomy. Depending on how units connect to each other, we can broadly distinguish between *feedforward* and *recurrent* networks. Whereas in a standard feedforward—i.e., *acyclic*—network, activations flow only in one direction (from lower to upper layers), in recurrent networks activations can flow in any direction at all. For ease of exposition, I employ a simple feedforward network in this section.

output units. Activations strictly feed forward. Each unit in layer *n* has output connections to each unit in layer *n+1*.

The input activation values of the input layer are propagated to the units of the next layer. Each receiving hidden unit will suffer a change in its level of activation. The resulting hidden activation values can be seen as a new activation pattern to be treated as a hidden vector in a n-dimensional space. Once again, the hidden units will feed their activation forward to the units of the following layer (either another hidden layer or the output one). In a similar way, an n-dimensional vector will be formed as a result of the new levels of excitation or inhibition of the n hidden or output units. Connectionist theorists hope that the interconnectivity among the units, and the flow of activation from one layer to another are sufficient to mimic and account for both the non-cognitive as well as the cognitive abilities of living creatures (see 4.5. below). When we try to explain, for instance, sensorimotor control, networks as simple as the feedforward model introduced above can be satisfactorily employed. A classic example in the literature is a crab that wants to grab an object situated at a certain distance in its visual field. The problem would be how to transform the *visual* information the crab is receiving into *motor* information that *tells* the crab *where* the object is with respect to its claw.[14] These networks operate by executing vector-to-vector transformations that allow the creature to go from a *sensorial* coordinate system to a *motor* one. Computations are thus better seen as vector-to-vector transformations from one coordinate system into

[14] For a comprehensive exposition of how neural networks can perform sensorimotor coordination tasks, see P.M. Churchland (1986) or P.S. Churchland (1986).

another. In short, we may appraise the dynamics of connectionist networks as the transformation of an activation vector (an input pattern of activation) into another vector (an output one) via one or several hidden vectors.

## 4.5 *Learning and Conceptual Organization in Neural Networks: State Space Semantics*

Connectionist networks as simple as the feedforward net shown above are very good at learning—see Rumelhart, McClelland et al. (1986), and Hanson and Burr (1990). Thanks to the employment of hidden units, multi-layered networks can develop *internal representations* that reflect the externally given inputs.[15] One learning technique extensively employed in connectionist modeling is the

[15] Very basic networks, such as Rosenblatt's (1959) *Perceptron* (a two-layered net with no hidden units) do employ learning rules—in particular the *delta rule*—that allows the network to alter its weights in order to reflect more appropriately environmental links between input and output patterns. The perceptron learns by readjusting the correlations between input and output units as a function of the deviation between each output unit's actual value and the expected value for that unit. Unfortunately, the lack of hidden layers, mediating between the encoded features at the input level and the output response, seriously undermined the learning capacities of the Perceptron, and soon, researches turned their attention to 'explicit-rules' learning machines—see 4.8 below. See Hanson and Burr (1990) for a comparison between Rosenblatt's Perceptron and the properties of more advanced networks with hidden units.

generalized delta rule (aka backpropagation).[16] To apply backpropagation, the network (initially activated with a set of random weight assignments) is fed with a particular input. As a result of a number of transformations, the network eventually produces an output vector. Since the network has been started with a random set of weights, it is highly probable that the resulting output vector does not coincide with the target pattern of output activation—i.e., the expected vector. Various of the output layer's units, then, must be in error. If we now take a single output unit and compare it with the target activation for that unit, we can take the difference in value as a *measure of error*. Then, by employing algorithms to adjust the weights, and by observing the effect that a minor positive or negative change in its weight would have in reducing the overall error, we can determine what kind of change (positive or negative) will make the output vector approximate more to the target pattern of activation. This process is then repeated for the rest of the connections from the upper to the lower levels in the network. What learning algorithms such as backpropagation thus try to accomplish is a minimization of the network's overall error by searching for a 'gradient-descent route' in the space of *potential weight assignments* (see below).[17]

Taking then a typical learning task such as categorization, a network trained

---

[16] Backpropagation was initially articulated by Rumelhart, Hinton and Williams as a generalization of the delta rule employed by Rosenblatt's Perceptron (see ft. 15)—see Rumelhart, Hinton and Williams (1986) for the *minutiae*.

[17] For a formal appraisal and illustrations of backpropagation see Bechtel and Abrahamsen (1991), and Elman *et al.* (1996).

by backpropagation can learn to classify fed stimuli into an established set of categories. A classical example in the connectionist literature is an acoustic network for sonar analysis. The network has a simple feedforward architecture—containing 34 input units, 14 hidden units and 2 output ones—and was trained by backpropagation to distinguish between rocks and mines.[18] In the training phase, the network is fed with the digitized outputs of an analyzer whose frequencies correspond to real sonar echoes bouncing back from both rocks and mines, lying at the bottom of the sea. The idea is that through backpropagation learning, the network can develop internal representations that reflect the externally given inputs. In the training phase, after each input has been fed, the weights are calculated upward. The result is that one of the two output nodes (one for the answer 'mine', the other for 'rock') will have a higher value than the other, eliciting thus an output response. If the network's response is correct—i.e., if the 'mine' node has a higher activation value when a mine-frequency has been given as input, for instance—, then the patterns of activation are left intact. Otherwise—if the answer is incorrect—the weights on the connections are recursively adjusted downward according to the measure of error calculated as a result of the difference between the actual response the network has given and what the correct answer should have been. Once the training phase is completed, the network is tested by feeding it with the frequency of several previously unencountered mines and rocks. The result Gorman and Sejnowski (1988) reported is that the network can respond *correctly*—

[18] See Gorman and Sejnowski (1988) for a detailed analysis of the training process, or Churchland (1989a) for a recapitulation of the key aspects of the network.

more than 90% success rate—to the novel inputs: If fed with a mine frequency, the network's mine output-node will have a higher level of activation. And similarly with respect to new rocks' sonar echoes. In this way, we may say that the network has *learned* to discriminate between mines and rocks.

Churchland makes use of Gorman and Sejnowski's results in order to articulate a connectionist-inspired theory of mental representation—aka *State Space Semantics*. The basic idea is that

> [the] brain represents various aspects of reality by a position in a suitable state space, and the brain performs computations on such representations by means of general coordinate transformations from one state space to another. (Churchland, 1986, p. 280)

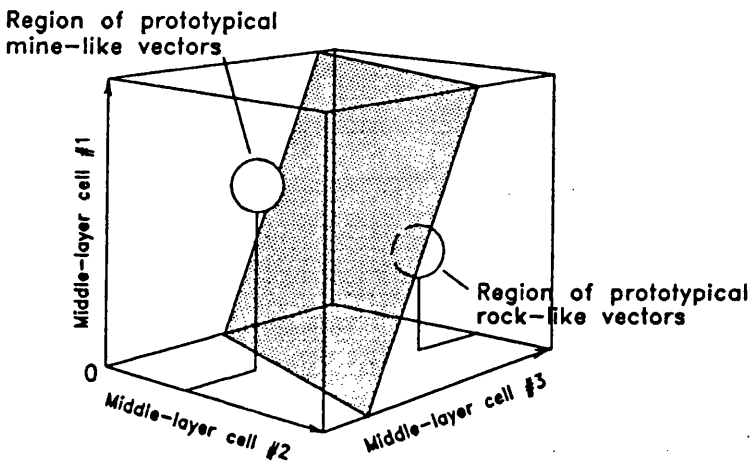
Churchland proposes that we understand *concepts* as points in a partial state space of a dynamical system. These points correspond to the tips of the vectors determined by the levels of activation of the different units in hidden layers—see 4.4 above. The semantic characteristics of a concept can then be seen as a function of the *place* that that concept—i.e., point—occupies in a geometrically characterized hyperspace. In this way, Churchland proposes, we may talk of semantic similarity between concepts in terms of the proximity of their respective absolute positions in state space, as identified in relation to a number of semantically relevant dimensions. In short, State Space Semantics tells us that the semantic connection of a concept $A$ with properties $x$ and $y$ can be analyzed in terms of the position of concept $A$ in a semantic space which is delimited in part by the $x$-

and *y*-dimensions.[19]

State Space Semantics sheds some light onto the ongoings of the mine/rock detector. By observing the behaviour of the hidden units during the training phase, and how the network performs subsequently under the presence of new exemplars, we find that the training process has *partitioned* the state space of possible patterns of activation across the hidden units. In particular, the internal space has been split into two sub-spaces: One corresponding to mine-representations, the other to rock-representations. One way to grasp this partitioning is via statistical techniques such as *Cluster Analysis*. It consists in pairing each pattern of activation with its closest neighbour. An average activation value is then calculated, and the process of pairing neighbours is repeated for the new patterns of activation. This technique is hierarchically applied, arriving in the end at a final clustering in space where points are located in several specific regions as a function of the similitudes shared with other points, culminating each sub-space in a *hot spot*, as Churchland calls it. The hot spot in the mine-like space is taken to represent the *prototypical* mine; the one in the rock-like space is taken to represent the *prototypical* rock (see figure 4.3 below). All the vectors whose activation values correspond to mine-inputs are seen as points in a mine-region, and the same goes for the 'rock' vectors. We may then judge how representative of a category an exemplar is as a function of the

---

[19] This is a simplification of Churchland's (1986) original proposal. Since then Churchland has developed State Space Semantics further, in order to meet a number of challenges. In particular, Churchland (1996; 1998) now defines semantic similarity, not in terms of *absolute* positions, but rather *relative* positions in state space—see chapter 5 below.




[Fig. 4.3]: The activation-vector space of the middle layer of the acoustic network for sonar analysis. Note the partition into two exclusive categories: mine echoes and rock echoes. Note also the two prototypical hot spots where typical and uncompromised examples of each category are routinely coded. (from Churchland, 1995, p. 83)

geometrical distance, in a pre-specified Euclidean space, that separates the vector in question from one particular hot spot. In this way, the task of the network can be better understood as a process of discrimination, not only between mine-like and rock-like representations, but also between more or less prototypical mines and rocks. Hence, under a novel input pattern, the network is discriminating the proximity of the new pattern of activation produced—represented as a point—to one or other of the hot spots. In short, the key point to bear in mind is that when a hot spot is activated, it represents the network's concurrent understanding of a given

environmental feature.[20]

As we saw above, connectionist learning consists in minimizing the network's error in the space of potential *weight* assignments. This statement however requires some further qualification that will help us appraise a crucial disanalogy between classical computers and neural networks as far as information storage goes. Even though neural networks can be trained to navigate complex domains, the information the network gathers from the environment is *not* explicitly stored anywhere, as is the case in a symbolic model. Digital computers store information in memory, such that relevant data for specific tasks can be retrieved and deployed when required. Connectionist models, on the other hand, don't retain information anywhere in the system. All the information present is the one being *actively* represented as an activation pattern at any given point in processing. When the network is not making use of a piece of information, that data is *nowhere* in the system. Nevertheless, if a network is to navigate a certain domain it must somehow

---

[20] It is noteworthy that Churchland's approach to concept organization finds empirical support in some psychological research on perception and categorization. Rosch (1975), for instance, challenged the rule-based perspective according to which falling into one category is determined by the satisfaction of an established—fixed—set of characteristics. In contrast, categorization is better understood in terms of prototypes. Think for example of *robins* and *ducks* as exemplars of the category 'bird'. Experimental psychology has taught us that the first exemplars fall more neatly into the 'bird' category than the latter. The reader can see that State Space Semantics can perfectly account for Rosch's dynamic understanding of categorization. On the other hand, Rosch's prototype perspective of concepts has been recently called into question. However, those against the Roschian

store information that can be necessary for future steps of processing. In connectionist learning it is the *weights*—i.e., the strengths—among the connections what gets stored. Ultimately, in the connectionist dynamical approach, we may say that the *knowledge* the network has of a target domain resides in the connection weights that have been generated during learning in accordance to a learning algorithm. The weights between the units is what allows the network to *recreate* all the patterns of activation corresponding to the features of the different stimulus in, for example, a categorization task (see 4.9 below). In this way, learning consists of certain weight adjustments such that the network comes to sort the given stimuli into the correct categories—i.e., such that a *single set of weights* allows the network to constantly generate the right activation patterns in the face of the activation from new input patterns. We now have an elementary picture of the basic components and dynamics of connectionist networks, and how concepts get organized in this framework. But before developing my connectionist defence of the Inscrutability Thesis in the next section, let me introduce a caveat regarding the *biological plausibility* of connectionist networks which is essential to avoid a source of misunderstandings concerning the potential application of connectionist modeling as an explanatory framework of human cognition.

Connectionist networks are *neurally* inspired. Research in neurobiology and cognitive science appears to favour connectionism as a fruitful model of cognition, over the computer metaphor advocated by classical artificial intelligence—see for

---

approach have produced models which are even more in accordance with Churchland's flexible construal of concepts—see, for example, Barsalou (1989), and Schyns and Rodet (1997).

example Churchland and Sejnowski (1992); and Elman *et al.* (1996). Among the reasons often cited, connectionist supporters stress certain tasks that neural networks appear to be better at than digital computers. Examples are pattern recognition, plan making, speech understanding, and in general, any task which is domain-specific.[21] Moreover, the friend of connectionism often alludes to two further features of connectionist networks which find clear neurobiological support. These are *real time* constraints on processing, and *graceful degradation.* Humans are able to perform highly complex tasks such as language processing in the order of hundreds of milliseconds. This imposes a serious limitation on processing since were we to perform complex tasks by following classical rules and programs, that couldn't be accomplished in more than 100 serial steps.[22] Real time constraints, connectionists argue, bring support to parallel processing. On the other hand, neural networks, like brains, seem to 'degrade gracefully'. Since information is distributed in parallel, optimal performance in a certain domain deteriorates gradually. On the other hand, classical processing is said to be 'brittle'. Either a system works or it doesn't, given certain damage that affects to a particular domain. So, in short, it seems that in some important aspects, connectionist networks exhibit properties of biological cognitive systems not found in classical symbolic models.

Granted that, nonetheless, it must be noted that connectionist networks bring a great deal of simplification in contrast with 'biological networks', failing to

---

[21] Digital computers, on the other hand, are far better at tasks requiring manipulation of large pools of data according to fixed explicit rules—e.g., number crunching.

[22] This has come to be known in the literature as the '100 step' constraint.

capture the fine-grained architectural and processing details of real brains. The following are important disanalogies, to name but a few: Firstly, the level of activation of connectionist nodes usually takes 2 values (in the case of *all-or-nothing* units), or $n$ values (if we discriminate between $n$ different activation levels between 0 and 1, where $n$ is not usually a very high number). In contrast, the spiking frequency of neurons ranges between 0 and 200 hertz. In this way, we might be able to distinguish a number $m$ of relevantly different levels of activation corresponding to values within this interval, where $m$ is considerably much higher than $n$. Secondly, a simple feedforward network has just one or maybe a few hidden layers, whereas natural networks can be formed by up to approximately 50 different hidden layers. Thirdly, as we saw, connections in feedforward neural networks are propagated from one layer to the immediate following one. In contrast, natural networks have *layers* connected with each other in a non-sequential order.[23] Moreover, we did not contemplate connections among the units of the same layer, whereas in natural networks, the level of activation of a given unit can be partially determined by its connections to other units of the same layer. And lastly, artificial networks operate with layers composed of hundreds, maybe a few thousands, of units. In contrast, biologically speaking, we can talk of layers with a number of units reaching into the millions! So, it seems that the dimensionality of the systems of coordinates to be determined by real brains is much higher than the

[23] There are many *recurrent* networks—as opposed to feedforward or non-acyclic ones—which do enjoy a richer non-sequential connectivity (see 4.6 below). Nevertheless, even the most developed recurrent networks are still far away from mirroring the connectivity patterns of brain cells.

dimensionality achieved by the hand of artificial neural networks. The number of potential positions of a vector in its relevant coordinate system will consequently be much bigger as well. This is important because *how* knowledge is stored may influence the course of future processing—see 4.7, and 4.8 below.

Furthermore, learning algorithms for weight-readjustment, such as backpropagation, depend crucially upon some form of *supervision.* The crucial point is that connection weights can only be adjusted by deriving a measure of error from a *target output.* This feature is not known to have biological implementation. Real nervous systems lack access to the target outputs that backpropagation exploits. And, moreover, even if brains did have access, information does not flow backwards in order to adjust weights as to lead to a better future performance of the system.

These considerations bring a substantial worry that must be addressed before I try to exploit connectionist theory for my defence of Quine's Inscrutability Thesis. Given that artificial neural networks are in a sense radically different from their biological counterparts, why should we pay any attention to connectionism in the first place if our subject matter are *real* cognitive agents? Putting it bluntly, a fair question is: What can connectionist theory possibly tell us about the brain? At this point, there is, I'm afraid, not one single answer that will satisfy everyone. Let me however sketch a couple of responses that will help to set to some extent the (modest) limits of the present work.

On the one hand, in fairness to the connectionist, it must be pointed out that modelers are progressively making use of more and more constraints under the light

of neurobiological research, refining thus their models so as to push the neural metaphor of connectionism as far as possible.[24] On the other hand, and more importantly, even though the richness of real brains cannot be easily implemented artificially, we may read connectionist theorizing at a more abstract level of understanding. Connectionist networks are to be interpreted as *abstract models* of real nervous systems. Even with simple abstract toy-models of this sort, the connectionist hope is that a lot about human cognition can be modelled. Connectionism can then set as its target the more modest project of accounting for the *coarse*-grained architectural and dynamical features responsible for human cognition. In this way, for example, even though we know that brains do not learn by backpropagation, the connectionist may still argue that *if* biological strategies to minimize error are *functionally* similar to backpropagation, then connectionist theory can still help us to understand the coarse mechanisms involved in human learning—cf. Elman *et al.* (1996).

My working hypothesis will thus be that these ideas, but perhaps with more complication, will equally shed light on higher-level cognitive problems—thought, language-mastery, etc. Ultimately, the best shot for my forthcoming arguments will

---

[24] New neural networks are being designed which can account, for example, for aspects of the mammalian visual cortex, human aphasia, etc. Some researchers are trying to find connectionist analogs of synaptogenesis, and synaptic pruning (see Elman *et al.*, 1996, p. 5; p. 49, and the references given there). Also, a number of strategies for implementing backpropagation with lower-level mechanisms have been pursued—see, for example, Hecht-Nielsen, 1989, and Parker, 1985; 1987. In addition, several other learning algorithms with increasing biological plausibility are being deployed.

be to frame them conditionally: If real brains do process thoughts and language via biological strategies functionally similar to the ones employed by recurrent neural networks—in particular, by making use of a non-classical form of constituency (see 4.6 below)—then Wright's 'psychological simplicity' argument will not go through. We now have the basic machinery to build up a connectionist defence of the Inscrutability Thesis.

## 4.6 *Simple Recurrent Networks and Conceptual Inclusion*

The processing of natural languages—where information is encoded serially—calls for the representation of complex hierarchical grammatical structures. A connectionist network that can master complex linguistic tasks must reflect the temporal dimension involved in language processing; an essential feature if we think for example of nested relative clauses, where grammatical *context* will determine the semantic properties of the words being processed. Connectionist networks proposed to date to account for this kind of complexity are, nonetheless, far from mimicking the complex patterns of human linguistic behaviour. Our interest, however, is to appraise *how* concepts may be represented in a connectionist architecture, even though these concepts will relate to a toy language—i.e., a small portion of a natural language. Recurrent networks are precisely designed to cope with the complex grammatical structures of the limited number of sentences of a toy language. The result is a non-classical approach to cognition where constituency and processing are *non-classical* in a sense yet to be explained.

A simple recurrent network is a standard feedforward net supplemented with one or more feedbackward pathways. The idea is to make use of this recurrent architecture in order to bring into play some sort of short-term memory. The information in state space at any given step of processing is fed back into the hidden layer of the network *along with* the 'normal' input pattern being fed at the subsequent step of processing. Thanks to this recurrence the network can process contextualized sequential information. Based on this recurrent architecture, Elman (1992)—see fig. 4.4—designed a network which does exhibit appropriate sensitivity to the syntactical dependencies found in grammatical structures.[25] Elman trained a recurrent network on a set of 10,000 grammatical sentences which were produced, in the classical rule-derived way, out of a lexicon of 8 nouns, 12 verbs, the relative pronoun 'who' and an end-of-sentence period. Items of this lexicon were randomly assigned a twenty-six bit vector. The input set consisted then of the successive concatenation of all the sentences in the pool of data formed out of the stream of these vectors. The network's task was to make correct predictions of subsequent words in the corpus of sentences. Being fed with a sequence of words from the input stream, the network had to predict the subsequent word. Using backpropagation—see 4.4—weights were adjusted to the desired output performance. Once the training phase was finished, Elman's network was tested on a set of novel sentences. As we shall see shortly, the prediction task for the net cannot be deterministic. Given a novel input, several correct outputs may follow.

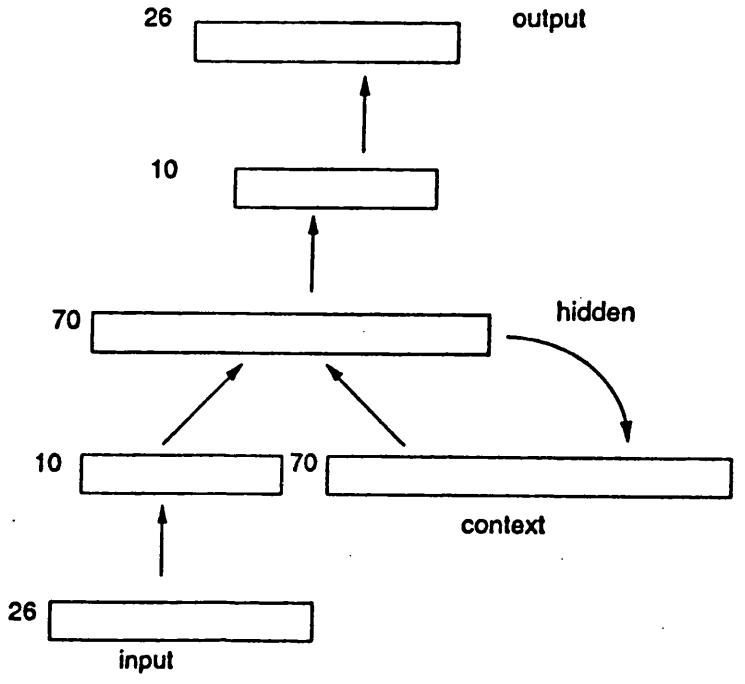
---

[25] The recurrent architecture employed by Elman is a variation of Jordan (1986) sequential network—see Elman (1988; 1989).

Probabilities of occurrence for all possibly correct predictions were determined by generating the likelihood vectors for every word in the novel corpus of sentences. The results were that the network could successfully discriminate grammatical strings of words.[26] The root mean squared (rms) error of predictions was 0.177 (sd: 0.463) against the likelihood vectors—for the details, see Elman, 1992, p. 154.

---

[26] Strictly speaking, the network's task is not to discriminate grammatically acceptable from grammatically unacceptable structures, but simply to make correct predictions of subsequent words—this conflation between the two tasks appears to be common in the literature. We may say, however, that the ungrammaticality of, for example, "boy who boys"—as a complete sentence—is indicated by the fact that the network does not predict a "." as a possible next word. That is, it recognizes that the sentence is not complete. If the string were "boys see boys" then the network would predict two kinds of possible next items: Namely, a period (which indicates that the sentence could be complete at this point); and also the word "who" (indicating that a grammatical continuation would involve a relative clause on the second noun). There are two reasons why we may want to stay with mere prediction. On the one hand, we may derive grammaticality from prediction by seeing whether the network believes that a sentence is (potentially) complete, or whether it wants additional input. In cases of degenerate input—e.g., "boys boys..."—the network predicts that nothing is possible as a successor. Thus, there are network behaviours which, although they do not explicitly indicate grammaticality *per se,* can be mapped onto grammaticality. Besides, we may model grammaticality explicitly by designing another network whose task is to examine Elman net's predictions, and output a 'grammaticality judgement'. On the other hand, prediction is a more ecologically plausible and naturalistic task than grammaticality. For present purposes, we need not expand on this issue, but just bear in mind that it is *not* a measure of grammaticality *per se* what Elman's network outputs. Many thanks to Jeff Elman for helping me clarify this issue.




[Fig. 4.4]: Elman's recurrent network used to discriminate grammatically correct sentences. (Elman, 1992, p. 153)

An illustration will help to appraise these results. Elman's net was presented with the following novel sentences, being fed one word at a time:

(a) boy who boys chase chases boy.
(b) boys who boys chase chase boy.[27]

Number information (e.g., boy/s) needs to be taken into account *over* the relative clause—who boys chase—common to (a) and (b). The results were encouraging.

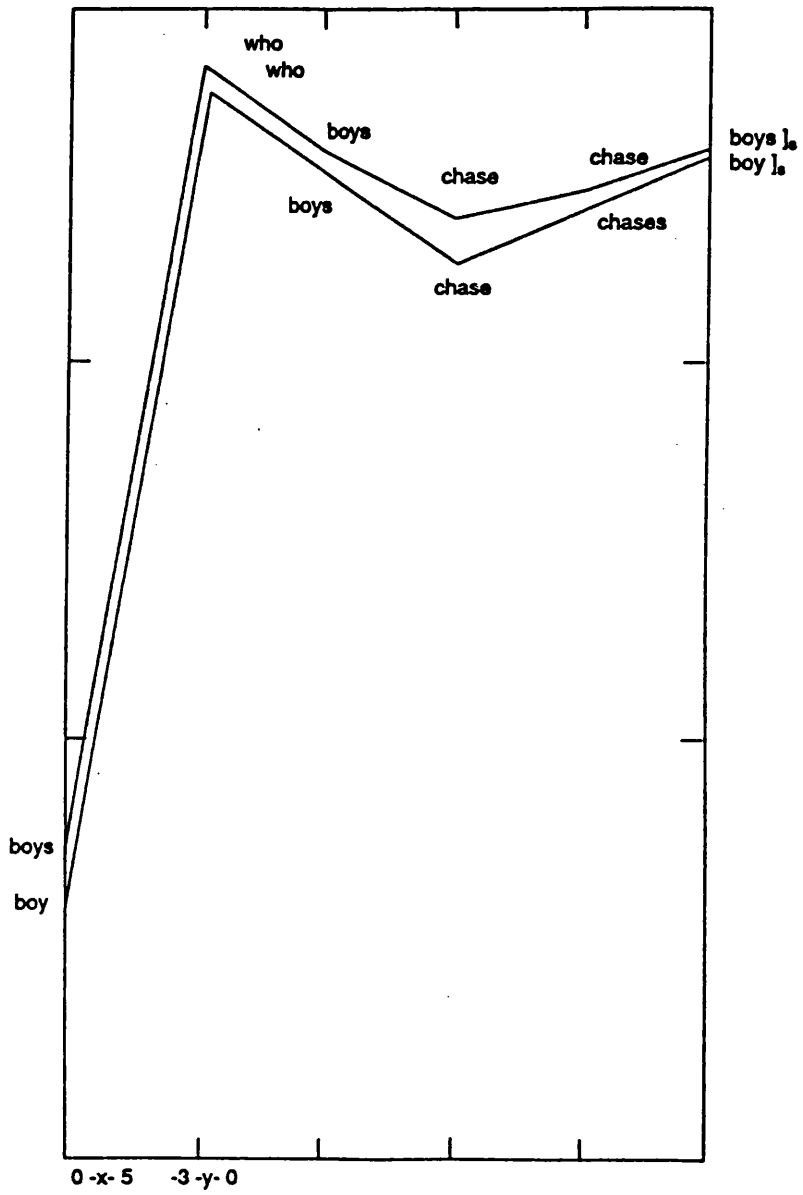
[27] English grammar demands: "boy whom boys chase chases boy", and "boy whom boys chase chase boy". We may nonetheless take sentences (a), and (b) above, for it doesn't make any difference to my forthcoming argument's.

Elman's net respected the grammatical agreement between the main clause subject and the main clause verb.[28] The crucial point for our purposes is to understand *how* Elman's network succeeds in its task. After each word in sentences (a) and (b) had been processed, the patterns of hidden unit activation were recorded. The hidden patterns of activation are distributed over 70 units, yielding a 70-dimensional state space. Making conceptual sense of the processing is thus not straightforward and requires some simplifying statistical treatment. We need to observe the temporal *trajectories* of these hidden patterns through state space. *Principal Components Analysis* (henceforth abbreviated PCA) provides us with a relatively simple way of looking into this high-dimensional sequential vector space. PCA is a dimensionality-reduction technique which consists in passing each member—sentence—of the input set through a trained network with its weights frozen, so that current learning does not interfere. The corresponding hidden patterns are then recorded and the number of *statistically relevant* correlations of the set of hidden activations is calculated. As a result, we get different vectors ordered by their values from greater to smaller amount of variance. These vectors recode each 70-dimensional input vector in terms of those variations, obtaining a more accessible—somewhat 'localized'—description of the hidden units activation patterns in which different vectors are used as first, second, ..., principal components in the analysis. If we now make use of the principal components—i.e., those input-output correlations that make the highest contribution to the net's overall output

---

[28] Similarly, Elman's net could represent successive embedding relationships, as found in complex relative clauses. See Elman, 1992, pp. 165-7, for the details.

behaviour—, we can see the temporal trajectories in the processing of sentences.




[Fig. 4.5]: Trajectories through state space for'[boy who boys chase chases boy' and 'boys who boys chase chase boy']. After the indicated word has been input, each point marks the

> position along the second principal component of hidden unit space. Magnitude of the second principal component is measured along the ordinate; time (i.e., order of words in sentence) is measured along the abscissa. [...The] sentence-final word is marked with a ]s. (Adapted from Elman, 1992, pp. 162-3)

Each different principal component carries different information. By examining the trajectories through state space along several dimensions when processing sentences (a) and (b), it was discovered that the second principal component played a key role in retaining number information of the main clause subject over the relative clause. PCA—see fig. 4.5 above—shows how

grammatically similar sentences, such as (a) and (b), follow closely resembling trajectories in the simplified space obtained by plotting the second principal component along the ordinate.

Let us now turn to a crucial philosophical implication of Elman's recurrent network.[29] Broadly speaking, the activations undergone by a connectionist network can be interpreted either *locally* or in a *distributed* fashion. In a localist model, individual units are used to represent entire concepts—see, for example, Rumelhart and McClelland (1982) model for word recognition. These localist units are atomic, and cannot thus be further decomposed. The semantics to be assigned in a localist interpretation are a function of the patterns of connectivity among these atomic units. By contrast, in a distributed model individual units are semantically uninterpretable. Representations are processed simultaneously by many units. Since a single ensemble of units can represent many concepts, we need to look at the entire pattern of activation in order to know which concept is being represented at a

particular time.[30] The representations obtained in Elman's net are *fully* distributed. The smallest interpretable pattern of activation is the one produced by the whole set of hidden units.[31] We cannot equate discrete parts of the hidden units' activation pattern with particular components of the sentences being processed. Nevertheless, Elman's network does capture the grammatical structure of the sentences it confronts. Sentences are *not* encoded by means of merely fully distributed *un*structured representations. Grammatical structure is reflected by coding grammatical variations as slight dynamical variations in the relevant activation patterns through state space. The syntactic contribution each word makes to the sentence is measured by the word's *own* level of activation, as encoded in hidden state space. The key issue, for our interests, is that connectionist and classical models differ in the way they represent constituency. Whereas in the classical symbolic approach, constituency is context-independent (see 4.3), connectionist

---

[29] For other philosophical implications of Elman's net see Ramsey, 1992, pp. 269-71.

[30] For our purposes we are interested in distributed representations. Since localist units are atomic, and connectionist semantics in these models develop as a function of the combination of these decomposable units, localist representations become functionally equivalent to the symbolic representations deployed in classical computationalism. Localist models would therefore provide manifestable evidence in favour of Wright's 'psychological simplicity' constraint—see 4.3 above.

[31] For some problems with fully distributed representations which I shall obviate, see Guttenplan (1994, pp. 203-4)

constituency is *context-dependent*.[32] Context-dependency gets encoded by the precise spatial location of a pattern of activation in representational space. Recent commentators have highlighted this crucial feature:

> [Elman's model] brings into play the idea of invoking different representations of the same concepts to capture certain structural relations. In this type of model, propositions do have individual concepts as constituent parts. However, this feature does not produce a straightforward implementation of LOT because of the way individual concepts are represented in such systems. In these models, the *form* of the representation of the concept itself—not its causal/functional relations with other concepts—determines its syntactic role in the proposition. In other words, we have implicitly 'stored' not one representation for a particular lexical concept but several different representations (encoded by patterns corresponding to different though nearby points in vector space), each of which account for a given syntactic role. Thus, we do not, on this picture, have a representation of BOY or APPLE but, rather, a cluster of representations of BOY-qua-[ ], APPLE-qua-[ ], where the bracketed blanks are filled in by the appropriate syntactic or conceptual role. (Ramsey, 1992, p. 269)

Given this connectionist perspective,[33] I contend, we find no motivation for

---

[32] For some classical examples of context-sensitivity see Smolensky's (1991) 'cup-with-coffee' story, and McClelland and Kawamoto (1986).

[33] Fodor's most powerful response to connectionism is that a connectionist model will not be able to explain the systematicity, productivity and inferential coherence of thought, unless it *implements* classical models, in which case LOT wins. An appraisal of Fodor's criticism would take us far afield. See Elman (1998) for a connectionist attempt to account for the systematicity of thought which avoids a symbolic implementation. I shall grant for the sake of the argument that a connectionist

Wright's 'psychological simplicity' argument. Imagine we feed Elman's recurrent network with several 'rabbit'-related sentences.[34] Take the following sentences:

(c) 'White rabbit', and
(d) 'White 99% undetached rabbit part'.

(c) and (d) are composed out of a set of lexically simple items: Namely, 'White', 'rabbit', '99%', 'undetached' and 'part'.[35] However, we should notice that in processing (c) and (d) the network does not store a *fixed* representation of the listed items, as classically identified. Rather, the network learns to use a cluster of representations of, say, 'rabbit'-qua-[syntactic/conceptual $role_1$], 'rabbit'-qua-[syntactic/conceptual $role_2$], where syntactic/conceptual $role_1$ is replacing f in 'White f' and syntactic/conceptual $role_2$ is replacing f in 'White 99% undetached f part'. The reason for this, as we've just seen, is the context-dependent character of the constituents. Each of the constituents to be distinguished in the structured sentences is encoded via different patterns of activation as a function of the context the constituent is embedded in. However, according to the above connectionist picture, there is no canonical representation of 'rabbit' to be singled out which is

---

architecture involves a connectionist model of cognition (see Rumelhart, McClelland *et al.*, 1986, chapter 4, p. 110). Nevertheless, I hope that by the end of the chapter (see 4.8 below) an idea of how to answer the charge of 'mere implementation' will begin to emerge.

[34] This is just a thought-experiment. I shall ignore the technical adjustments required in the architecture and training regime with respect to Elman's above simulation.

[35] For economy, we may ignore that '99%' can be further decomposed into the following lexical items: '99', 'per', and 'cent'. Similarly for 'undetached'.

common to 'rabbit'-qua-[syntactic/conceptual $role_1$] and 'rabbit'-qua-[syntactic/conceptual $role_2$] . Instead, there are two different representations encoding for *each* different sentential context. Were we to apply PCA on the sentences 'White rabbit' and 'White 99% undetached rabbit part', we would find that both sentences would follow different, although somewhat resemblant, trajectories in state space. 'rabbit'-qua-[syntactic/conceptual $role_1$] and 'rabbit'-qua-[syntactic/conceptual $role_2$] would occupy different positions reflecting thus different paths through space as a function of the previous words being processed. Obviously, those 'rabbit'-related vectors representing a similar grammatical role will tend to gather in certain subregions. However, the net performs its task at the level of the numerous context-dependent and distributed internal states. In this way, we should not see the idiosyncratic representations of 'rabbit' as word-tokens of the same word-type, as LOT and its classical form of constituency maintain. Whereas under the LOT hypotheses—see section 4.3—exercising the concept expressed by 'rabbit' was the tokening of the corresponding expression of LOT—viz., RABBIT—, my working hypotheses is that the conceptual repertoire expressed by 'rabbit' in an utterance of 'White rabbit' is whatever real internal state the connectionist theory maps 'rabbit' onto. Likewise, the conceptual repertoire expressed by 'rabbit' in an utterance of 'White 99% undetached rabbit part' is whatever real internal state (the same or different) the theory maps that utterance of 'rabbit' onto.

Assuming this connectionist setting, Wright's argument against the Inscrutability Thesis loses its grip. It is not the case that 99%-UNDETACHED-

*RABBIT*-PART includes a constituent RABBIT which allegedly is *common* to other RABBIT-related representations. The constituent RABBIT in WHITE RABBIT is a token of a *different* type from the constituent RABBIT in WHITE 99%-URP. Lexical inclusion in English, hence, does not imply conceptual inclusion. So, when we are confronted with several contextualized, though semantically related, concepts, we should conclude that none of them includes the others. In this way, neither the phrasal concept 99%-URP includes the lexical concept RABBIT, nor the other way round.[36]

I conclude then that the Quinean can go with modern scientific fashion and make use of the '99%-urp' referential scheme. A hard-line Quinean will ignore Wright's argument if it appeals to unashamedly mentalistic concepts. On the other hand, by approaching the issue of the naturalization of concepts in a way acceptable to a Quinean—see 4.2—we have at least two options: either we identify concepts with the orthodox classical features championed by Fodor under the LOT hypothesis, in which case Wright may hold to his principle of psychological simplicity, or we identify concepts with non-classical features acceptable to a connectionist, and then, as we've seen, Wright's argument does not go through.

The above discussion, together with the arguments offered in chapter 3,

---

[36] It is worth remarking that on the connectionist view the finite set of basic clauses of a translational manual does *not* give a basic repertoire of concepts from which all other concepts are constructed. The connectionist basic-to-phrasal direction of conceptual formation is orthogonal to the requisites imposed by LOT's classical constituency. The basic clauses are lexically basic, but have no privileged conceptual status—see 4.8 below.

constitute the bulk of my defence of Quine's Inscrutability Thesis. In the remainder of this chapter, and chapter 5, I shall try to address three important criticisms launched from the connectionist corner that may well torpedo my whole project.

## 4.7 *Statistical Analyses, Symbolic Approximation, and Causal Efficacy*

In this section I shall consider two routes the anti-Quinean may pursue in order to reply to the above connectionist defence of the Inscrutability Thesis. The general idea underlying both rejoinders is that statistical techniques extensively employed in data analysis—statistical techniques such as Cluster Analysis (see 4.5) and PCA (see 4.6)—may reveal that the connectionist approach to cognition, and in particular the connectionist approach to language processing, does not differ substantially from the symbolic approach reviewed in section 4.3. To bring up the core idea, the anti-Quinean may argue that connectionist constituency—i.e., context-*dependent* constituency (4.6)—can be 'statistically forced' into a classical—i.e., context-*free* (4.3)—mould. Were this to be the case, the anti-Quinean will contend, the results achieved in section 4.3, where classical constituency favoured a standard interpretation of Native, may equally apply once we adopt a connectionist architecture. To make a long story short, we may say that statistical techniques can help to close the gap between connectionist and classical models of cognition. This 'symbolic approximation' (see below) may be accomplished in at least two ways, giving rise to two possible lines of argument for the anti-Quinean to exploit. In what follows I shall flesh out both rebuttals. Then I shall suggest two reasons as to

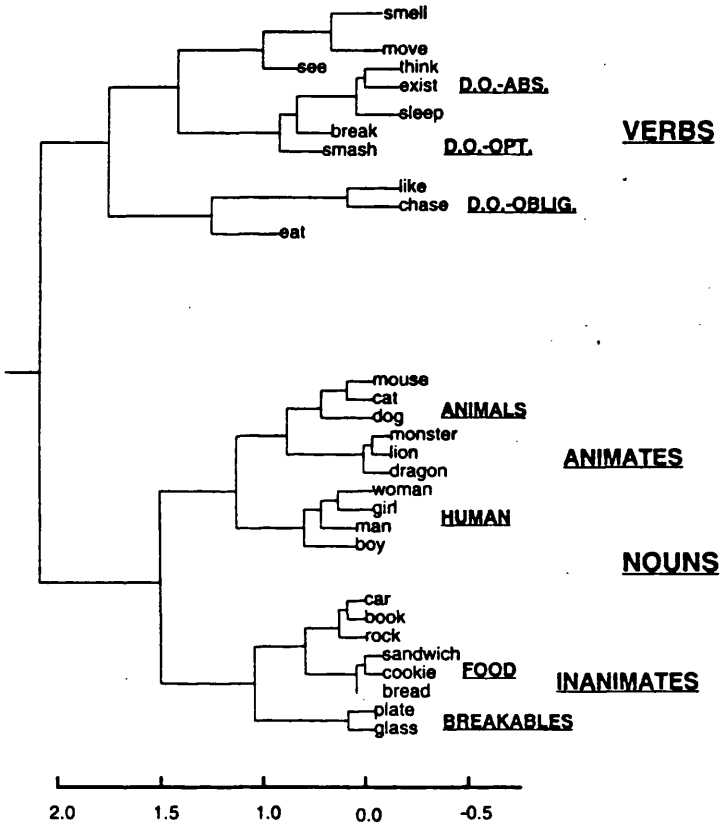
why the anti-Quinean cannot help herself two either of the two lines of response in order to reply to the connectionist Quinean of section 4.6.

*First Anti-Quinean Rejoinder* A lesson to learn from the employment of distributed neural networks—where single hidden units have no representational power in isolation (see 4.6 above)—is that we must drop altogether the classical decompositional approach to cognition, and adopt a mathematically-inspired one where the underlying mechanisms of cognition are explained statistically. Statistical techniques, such as cluster analysis provide us with a higher-level description of the piecemeal dynamics of the units-and-weights computations of the neural level. These higher-level analyses permit connectionist representations to capture, for example, information about abstract lexical categories of the sort typically referred to by classical symbols. A cluster analysis of a simple sentence-processing simulation run by Elman (1990) will serve to illustrate this crucial feature.

Elman (1990) trained a simple recurrent network, containing 150 hidden units, to predict successive words in sequences of sentences.[37] After the network learned the appropriate input/output mapping, Elman performed a hierarchical clustering of the 150-dimensional hidden activation space. As we saw when interpreting the behaviour of the acoustic network for sonar analysis (section 4.5 above), by pairing together all those inner states as a function of the spatial

[37] For present purposes we need not dwell on the details of Elman's simulation. The reader may care to consult Elman (1990) for the *minutiae*, or Elman *et al.* (1996, chapter 2) for a recapitulation of

proximity among the diverse activation patterns, cluster analysis enables us to visualize indirectly high-dimensional spaces. Figure 4.6 shows the spatial organization of the hidden representations generated by Elman's network.




[Fig. 4.6]: Hierarchical clustering of hidden unit activation patterns from the sentence-prediction task (Elman, 1990). The network learns distributed representations for each word which reflects its similarity to other words. Words and groups of words which are similar are close in activation space, and close in position in the tree. (Elman *et al.*, 1996, p. 96)

---

some key aspects of the network. Although the tasks are similar, the reader should not mistake this simulation for the one reported in section 4.6.

As the hierarchical clustering of figure 4.6 illustrates, the representations created in Elman's network reflect, at a high-level of description, certain abstract categories. In particular, we can see that the tree structure of the words' patterns of activation has divided the hidden space into two groups corresponding to the categories NOUN and VERB. These two categories comprise the 29 lexical items learned by the network. If we move from left to right in the diagram we can see, for example, that 'boy' relates to an activation pattern belonging to the category HUMAN, which in turn is contained within a broader region identified as ANIMATES, which ultimately corresponds to the higher-level category NOUN. By looking at hidden spatial proximities in this way, the connectionist has a way of framing the *type/token* distinction which may furnish the anti-Quinean with the perfect tool for her purposes. The crucial idea is that, similarly to our distinction in spatial terms between a noun and a verb category, we may say that tokens that belong to the same representational type are neighbours, or more precisely, are in closer proximity to each other than to tokens belonging to other types. Let me flesh out this distinction in order to pave the way for the first anti-Quinean rejoinder.

To see the threat that cluster analysis poses to the Quinean, we need to move to a deeper level of analysis in Elman's above simulation. For reasons of computational economy, Elman (1990) performed a cluster analysis of the hidden activation patterns *once* those patterns had been *averaged* over many different contexts. Hence, the hierarchical clustering of figure 4.6 picks out internal representations of lexical items which are *context-insensitive*. So, for instance, the activation pattern taken to represent 'boy' was the *mean* vector obtained by

averaging out many idiosyncratic presentations of 'boy'. To gain more accuracy, Elman did a second cluster analysis; this time, of the *real* context-dependent activation patterns. The goal was to observe whether the tree structure of figure 4.6 reflected what the network actually learned, or whether it was an artifact created by the employment of mean—context-free—vectors. Since the network developed 27,454 hidden activation patterns, it would have been difficult to display graphically the tree structure obtainable, had we performed a hierarchical cluster analysis. However, we can appraise intuitively what's going on. The tree structure, Elman reports, is similar to the one graphed in figure 4.6, with the exception that each branch is further arborized in order to reflect specific contexts within each lexical item. Elman's following comment brings to the fore the moot point for our current concern:

> It would be correct to think of the tree in [figure 4.6] as showing that the network has discovered that there are 29 *types* (among the sequence of 27,454 inputs). These types are the different lexical items shown in that figure. A finer grained analysis reveals that the network also distinguishes between the specific occurrences of each lexical item, that is, the *tokens*. The internal representations of the various tokens of a lexical type are very similar. Hence, they are all gathered under a single branch in the tree. (Elman, 1990, p. 205)

The key question implicitly addressed in the above quote is: How can we represent the type/token distinction in distributed connectionist networks? Figure 4.6 revealed how abstract lexical categories such as NOUN and VERB could be set apart in terms of the subregions of hidden space marked off by looking at different

degrees of spatial proximity. In like vein, we may say that the type/token distinction gets cashed out by the statistical development of boundary regions in hidden space. It is the fact that different context-dependent tokens fall within the boundaries of a given subregion what makes them belong to the same representational type.[38]

At this point, a clear objection breaks through.[39] Consider again the recurrent network of section 4.6 being hypothetically fed with several 'rabbit'-related sentences. As I argued earlier (see 4.6 above), theoretically we should expect the network to represent several 'rabbit'-related sentences by different positions in hidden space; different positions reflecting the sentential context of each particular occurrence of 'rabbit'. Given that, I contended, the constituent RABBIT in for example WHITE RABBIT is a token of a different type from the constituent

---

[38] This appears to be the orthodox reading of the type/token distinction in the connectionist literature (see, for example, Miikkulainen, 1993). In my opinion, it would be more fruitful to formulate a rather more radical proposal, according to which the process of differentiation of tokens belonging to one representational type—i.e., falling within the boundaries of a subregion in hidden space—turns out to be equivalent to the elaboration of *different types altogether*. After all, given that tokens (allegedly belonging to one type) preserve contextual idiosyncrasies, it would be tempting to argue that, granting connectionism, the type/token dichotomy becomes an artifactual distinction. Tokens, we may venture to say, become types themselves. Elaborating on this thought, nevertheless, would take us far afield. I shall follow for present purposes the orthodox—spatially inspired—connectionist portrayal of the type/token distinction.

[39] The forthcoming objection should be read as a criticism launched by a hypothetical connectionist foe of Quine. To the best of my knowledge there is no argument in the connectionist literature explicitly profiting from cluster analysis to argue against a connectionist version of Quine's Inscrutability Thesis.

RABBIT in WHITE 99% UNDETACHED RABBIT PART. Unfortunately, the preceding discussion puts in plain sight that my conclusion was at least premature. Thanks to statistical techniques such as cluster analysis, we have a scientific level of description of speakers' brain states such that we can group together all those states involved in inferences depending on the lexical item RABBIT. Elman's above clustering highlights the fact that all those 'rabbit'-related vectors representing a similar grammatical role would tend to gather in certain subregions. In short, the family resemblance of the different contextualized representations of anything rabbity can be grouped together culminating in a statistical *unity*. I shall call this central tendency RABBIT*. Now, the fact that connectionism treats constituents in a non-classical, context-dependent, way need not cause any concern to the anti-Quinean. The reason is that above the connectionist fine-grained level, there is a higher statistical level of understanding which provides us with the conceptual stability the anti-Quinean requires. In this way, someone may contend, the connectionist defence of the Inscrutability Thesis offered in section 4.6 does not go through. The reason is simply that the constituent RABBIT in WHITE RABBIT and the constituent RABBIT in WHITE 99%-URP are, not tokens of different types as I argued earlier, but rather tokens of the same type: namely, the type RABBIT*. All different 'rabbity' constituents can be seen as spatially grouped together under the communal concept RABBIT* which is stable enough to play the role that the classical constituent RABBIT played in the Fodorian model which favoured a standard interpretation of Native (see 4.3).

Cluster analysis nicely illustrates spatially how abstractions can emerge

statistically from the fine-grained dynamics of neural nets. However, the point is more general, and could be made with virtually any other statistical technique. After all, the role of statistical analyses is precisely to reduce dynamical detail so as to have a firmer (symbolic?) grip of what's going on at the nuts-and-bolts level of processing. Since the second anti-Quinean rejoinder runs along similar lines—i.e., trying to close the gap between classical and connectionist models of cognition—let me briefly sketch it by looking at the principal components analysis performed on the simulation run by Elman that we reviewed in section 4.6.

*Second Anti-Quinean Rejoinder* The anti-Quinean's misgivings, on the other hand, may be confined to the fact that the orthodox interpretation of Elman's net that we find in the secondary literature (e.g., Ramsey, 1992, —see 4.6 above) is wrong. That interpretation is that constituency can be kept context-dependent by encoding the precise location of each individual (idiosyncratic) pattern of activation in state space. The failure, the objection would run,[40] is the result of thinking about the mathematics and statistics of what's happening in the representational space of Elman's net in a muddled way: Coding grammatical variations as slight different positions in hidden space does *not* imply that Elman's model is not subject to a *classical symbolic* treatment—i.e, does not imply that there are not context-*independent* representations. Notice that were we to do a PCA on a classical model (by treating different registers as dimensions, for example), we would get results

[40] This worry was raised by an anonymous referee of *Mind and Language* in response to a previous version of this chapter—see Calvo Garzón (2000a).

similar to those obtained on Elman's network. Grammatical variations between, say, a-as-object and a-as-subject would be reflected by slightly different positions in representational space. Similarly for connectionist models, the fact that on a PCA, a-as-object and a-as-subject are located in different positions does not mean that there is not a perfectly good context free symbol. We might obtain context-independent representations in Elman's model by paying attention to *particular* principal components in the analysis. Specifically, activation of the first principal component would be a good candidate since, as we saw in section 4.6, the contextual difference between the sentences 'boy who boys chase chases boys' and 'boys who boys chase chase boys' (see figure 4.5 above) seems to be captured by location on the *second* principal component. Now, once we interpret Elman's model as dealing with a classical form of constituency, Wright's 'psychological-simplicity' argument may kick in for the reasons rehearsed above, discrediting thus Quinean perverse alternatives to our standard semantic theory.[41]

Moreover, someone may try to close the gap between classical and connectionist interpretations by permitting 'infinite precision analog' between classical and connectionist models. In that case, it is indeed not clear that the representational capacities of a classical symbolic model can be distinguished from

---

[41] I believe, nonetheless, that the burden of proof is still on the proponent of this second rejoinder. Notice that the fact that the second principal component captures contextual idiosyncracies is logically consistent with *other* principal components capturing further subtle differences, or reflecting a different contextual role altogether. However, this point is too technical to be of import for the purposes of this chapter.

the representational capacities of a recurrent network. However, such a move would give up precisely what makes symbols attractive. Namely, their ability to abstract over classes of items, *ignoring context*. It is unlikely, nevertheless, that the anti-Quinean would be willing to pursue this last line of argument, and I shall not press on it further.

Summing up, both rejoinders exploit the use of statistical techniques to bridge the gap between classical and connectionist approaches to constituency. In the first case, context free symbolization emerges as a result of the statistical clustering of different, but semantically related, representations. In the latter case, the anti-Quinean hopes that particular principal components in a PCA will symbolize the network's internal representations. In the remainder of this section, I shall offer two considerations aiming to disprove the anti-Quinean's use of statistical analyses for her purposes. In a nutshell, the two reasons are that the anti-Quinean is ignoring: (i) that statistical abstractions loose processing detail, and (ii) that the abstractions generated statistically are *causally inert*, and thus cannot play any explanatory role as far as the dynamical processing of connectionist networks is concerned. (i) bears directly on the second anti-Quinean rejoinder. On the other hand, (ii) is a more general criticism, and aims to target both rejoinders. Let me expand on these two points.

**(i)** Using PCA to 'symbolize' the hidden unit representations might in fact give the appearance that connectionist representations are equivalent to symbols. Unfortunately, this move would fail to completely reflect important variations in

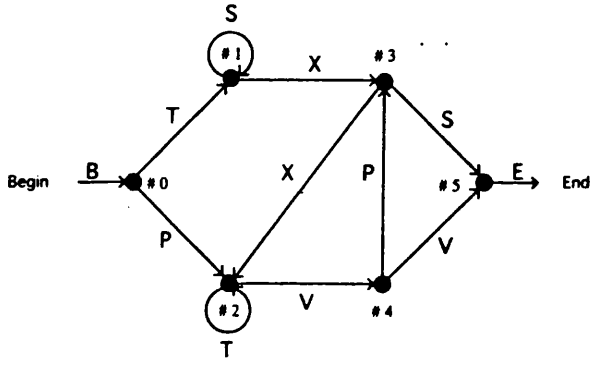
form that are not captured by PCA. That is, such a move gives the illusion that the network trades in symbols, when in fact there are important subtle variations in hidden unit representations; variations which may have causal consequences for the network's behaviour. The general point, to put it bluntly, is that statistically-generated abstractions loose computational detail. Statistical techniques such as PCA are meant to preserve remaining regularities across contexts. In one sense, then, symbolic models approximate connectionist ones. However, the approximation works only to a certain extent, and the gap between classical and connectionist accounts of constituency, I contend, will ultimately never be fully bridged. Commenting on the recurrent network I made use of in section 4.6, Elman notes:

> The fact that the networks here exhibited behavior which was highly regular was *not* because they learned to be context-insensitive. [... Even when] these networks' behavior seems to ignore context the internal representations reveal that *contextual information is still retained.* (Elman, 1991, p. 220; emphasis added)

Elman's remarks capture the essence of my first counter-argument. To expand on it, I shall next report on two well-known neurosimulations in the connectionist literature run by Servan-Schreiber, Cleeremans, and McClelland (1988).

Servan-Schreiber *et al.* designed a network for processing sequences of letters randomly generated by a finite-state grammar.[42] The general goal was to appraise *what sort of representations* simple recurrent networks make use of. Similarly to

Elman's net (see 4.6 above), the present task is to predict successive letters in a sequence. Sequences of letters are generated by Reber's (1967) finite-state grammar. A finite-state grammar is a closed circuit made up of nodes which are connected by arcs (see fig. 4.7). The nodes stand for possible states the system can be in.




[Fig. 4.7]: The finite state grammar deployed by Reber (1967; 1976). Numerals indicate states, and letters indicate transition arcs. Sentences are generated by traversing a path from initial state #0 to final state #5. After each transition a letter is produced, obtaining thus sequentially a string. (From Servan-Schreiber *et al.*, 1988, p. 6)

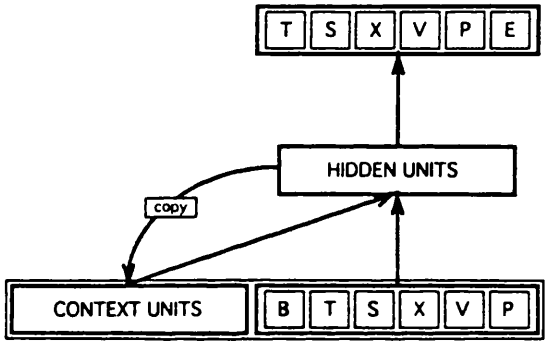
Grammatically correct sequences are produced by moving through the circuit from the 'start' node to the 'end' node. The different paths between 'start' and 'end' correspond to the several possible transitions from one intermediate node to another.[43] A probability of .5 is granted for every possible transition. In this way, we obtain a number of grammatical sequences by following different paths from

---

[42] For an extended treatment of this and related issues see Cleeremans (1993).

[43] The reader should not confuse the finite-state grammar circuit with common neural network architecture. Although the diagrams are somewhat resemblant, the above circuit is meant to be implemented by a *serial* machine such that at any given time only *one* state can be activated.

node to node. So, Reber's grammar will generate sequences such as, for example, TXS, PVV, TSXS, or PTVPS. In the prediction task set by Servan-Schreiber *et al.* the target for the network was to predict successive letters in a string being fed to the network letter by letter.[44] A network based on Elman's recurrent architecture was used to achieve this task (see fig. 4.8).

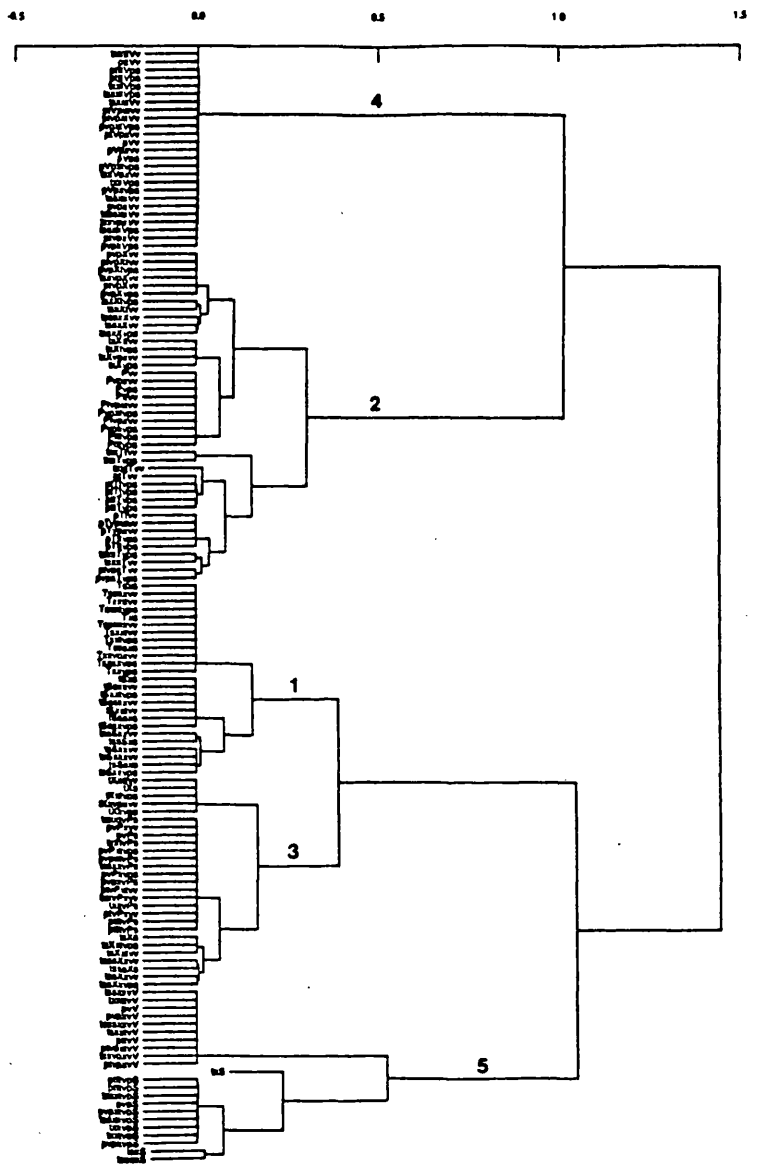



[Fig. 4.8]: General architecture of the network. (from Servan-Schreiber *et al.*, p. 7)

Servan-Schreiber *et al.* trained several networks based on this recurrent architecture. In an initial experiment they trained a network with just three hidden

[44] Since given a sequence of letters several possibilities may follow, the network needs information about the path traversed (not simply about the preceding letter). The challenge is similar to the one reviewed in section 4.6, where Elman's net needed to take into account number information over relative clauses. In like vein, arcs bridging nodes in Reber's finite-state grammar may be labeled with words, rather than letters. Hence, TXS could be interpreted as 'boys like girls'. Generally speaking, Elman and Servan-Schreiber *et al.* are faced with the same problem. Namely, to explain how networks can reflect the temporal dimension inherent in mastering increasingly complex linguistic structures.

units. The network was trained on a base data set composed of 200,000 strings




[Fig. 4.9]: Hierarchical Cluster Analysis of the H.U. activation patterns after 200,000 presentations from strings generated at random according to the Reber grammar (Three hidden units). (From Servan-Schreiber *et al.*, p. 10)

generated by Reber's grammar. After the network reached a successful degree

of performance, Servan-Schreiber *et al.* performed a hierarchical cluster analysis of the hidden units patterns of activation (see fig. 4.9):[45]

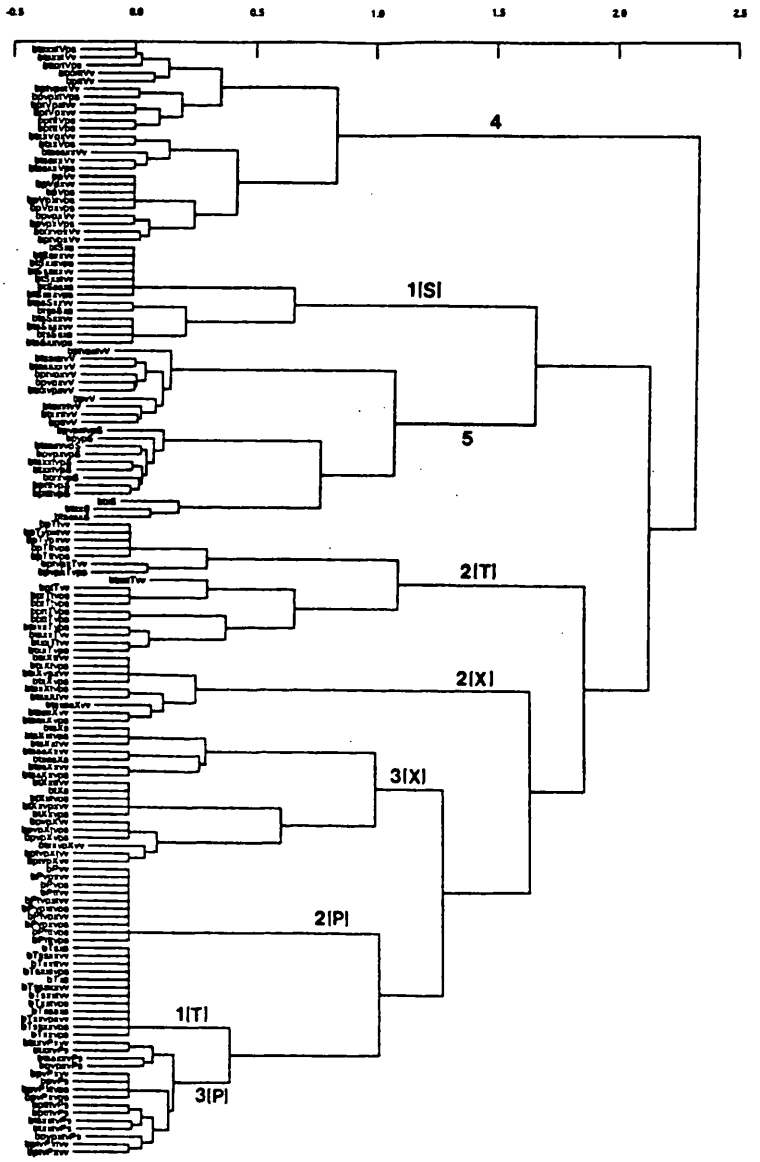
Figure 4.9 shows a tree structure where strings being represented in hidden space similarly are clustered together. Paying close attention to the clustering, we can see that hidden patterns of activation are grouped separately for each different node in Reber's grammar. At the bottom of the tree, for example, we can see that all patterns that activate the 'end' node are clustered together. From these results, Servan-Schreiber *et al.* conclude that the recurrent network behaves in accordance—to a degree of approximation (see below)—to the theoretical finite-state automaton corresponding to Reber's grammar. This clustering profile seems to encourage at first sight the anti-Quinean position according to which classical and connectionist models are not as different as initially thought to be.

Things, however, are not that simple. We can see why by looking at a different simulation. Servan-Schreiber *et al.* trained another recurrent network on the same prediction task, but this time, instead of three, the network had 15 hidden

---

[45] For the training details see Servan-Schreiber *et al.* (1988), pp. 6-11. Although the target of my first response is the 'symbolic emergence' allegedly achieved via PCA, for purposes of illustration I report on a cluster analysis study. A shortcoming of cluster analysis is that it preserves only *spatial* information, loosing the *temporal* dimension involved in processing sequential inputs. PCA, on the other hand, tells us what the trajectories between hidden states at different stages of processing look like (see 4.6). For present purposes, we need only focus on how cluster analysis risks overlooking representational idiosyncracies—bearing in mind that a similar point, although more complicated, could be made with respect to PCA.

units. Figure 4.10 below shows the cluster analysis performed on the trained




[Fig.4.10]: Hierarchical Cluster Analysis of the H.U. activation patterns after 200,000 presentations from strings generated at random according to the Reber grammar (Fifteen hidden units). (From Servan-Schreiber *et al.*, p. 14)

network. If we compare the clustering previously obtained in the three-

hidden-units network with the clustering obtained in the fifteen-hidden-units network, we can see that the latter (the higher-dimensional network) delivers a much more complex clustering of the hidden activation patterns. Broadly speaking, there are more levels of hierarchical clustering, and more subclusters within levels. By paying close attention to the hierarchical clustering in figure 4.10, the reader can see that nodes appear divided, retaining information about the particular arcs leading into them.[46] If we look further left in the tree structure, we can see that highly detailed information about paths leading to letters is retained. So, for example, sequences of the form T*VPS* and X*VPS* are located in *different* subclusters.

Servan-Schreiber *et al.* call our attention to the reasons for this increase of representational complexity:

> [We] should point out that the close correspondence between representations and function obtained for the recurrent network with three hidden units is rather the exception than the rule. With only three hidden units, representational resources are so scarce that back-propagation forces the network to develop representations that yield a prediction on the basis of the current node alone, ignoring contributions from the path. This situation

[46] To keep the record straight, only nodes #1, #2, and #3 show such 'redundancies'. Nodes #4 and #5 seem to ignore contribution from the arcs feeding into them (see figure 4.10). This, however, may be a byproduct of the architecture chosen. Plausibly, more hidden units would yield a richer subclustering. For present purposes, focusing on nodes #1, #2, and #3 is enough to drive my point home.

> precludes the development of different—redundant—representations for a particular node that typically occurs with higher numbers of hidden units. When redundant representations do develop, the network's behavior still converges to the theoretical finite state automaton—in the sense that it can still be used as a perfect finite state recognizer for strings generated from the corresponding grammar—*but internal representations do not correspond to that idealization.* (Servan-Schreiber *et al.*, 1988, p. 12; emphasis added)

We can now see why the anti-Quinean second rejoinder is doomed to failure. The worry was that particular principal components in a PCA could deliver context-independent—i.e., symbolic—representations. To remind the reader, the reason offered was that contextual idiosyncrasies were captured by location on the second principal component. The anti-Quinean then hoped that the *first* principal component would make Elman's network subject to a symbolic treatment. The research carried out by Servan-Schreiber *et al.* clearly illustrates—although see fn. 44 above—why the anti-Quinean is being misguided. The claim that 'coding grammatical variations as slight different positions in space is compatible with treating Elman's model symbolically' is correct only in a loose sense. The 'symbolic approximation' works to the extent that the behaviour of Elman's network converges to the function it's being trained to perform by backpropagation. Namely, to predict subsequent words in sequences of sentences. Elman's network does master abstract generalizations. However, this does not give us the full picture of the representational capacities of connectionist networks. Even though the behaviour driven by backpropagation can be said to deliver a symbolic approximation, hidden units will tend to gather as much computational detail as

possible; detail not strictly required for the network in order to make accurate predictions. In this way, even though contextual divergencies in Elman's net were captured by the second principal component, and even assuming for argument's sake (although see fn. 40) that other principal components may portray context-independent representations of the sort required by classical models, the anti-Quinean rejoinder still wouldn't be sound. The reason is simply that the representations obtained in hidden space still retain context-dependent detail—even though backpropagation only 'guides' the network to capture an abstract functional mapping. As Elman points out:

> [Connectionist models can] lead quite naturally to generalizations at a high level of abstraction where appropriate, but the behavior remains ever-rooted in representations which are contextually grounded. (Elman, 1991, p. 221)

On the other hand, as Servan-Schreiber *et al.* point out, lack of representational resources—i.e., scarce hidden dimensionality—leads to abstract, context-free, representations. That is the best (worst?) case scenario where a symbolic approximation may be accomplished. However, the more representational resources at hand, the less accurate the approximation between classical and connectionist models will be.[47] In short, the outcome of reading statistically-

---

[47] That symbolic models can merely approximate connectionist models has been repeatedly stressed in the connectionist literature. Notoriously, Smolensky (1988) highlights the cleavage between symbolic and subsymbolic levels of analysis. This distinction will have important implications in the debate over the need for symbolic representations and rules as urged by Fodor and Pylyshyn (1988), and Pinker and Prince (1988)—see 4.8 below.

generated symbolic representations too seriously is a failure to explain the actual mechanisms of cognition.[48]

The above considerations, and in particular the lack of a robust correspondence between representations and function in connectionist networks, drive us to the second reply which can be seen as the other side of the same coin.

**(ii)** The moral of Servan-Schreiber *et al.*'s simulations is that connectionist representations are too rich to be identified with classical abstractions of the sort generated statistically in connectionist theory. The other side of the same coin is that the abstractions generated statistically are causally inert, failing thus to play any explanatory role whatsoever as far as the dynamics of connectionist networks is taken as our model of cognition—see 4.8 below. Let me elaborate.

The gist of the first anti-Quinean rejoinder was that high-level abstractions generated by cluster analysis (see above) correspond to the folk psychological concepts posited by classical models. In this way, the statistical central tendency of a pool of 'rabbit'-related inputs—viz, RABBIT*—corresponds to the context-independent concept that those sympathetic to the Fodorian classical approach call RABBIT. There is however, I contend, a crucial disanalogy between the RABBIT* that emerges via cluster analysis, and the RABBIT of Fodor's LOT. Namely, that whereas RABBIT* is a mere abstraction, and is thus causally inefficacious,

---

[48] It is noteworthy that symbolic approximations may prompt the associated risk of missing potentially relevant information for future stages of processing. Fleshing out this issue, however, would take us far afield.

RABBIT (under Fodor's reading!) is causally efficacious.[49] The anti-Quinean first rejoinder, I believe, is the result of overlooking this key disanalogy between connectionist and classical models.

Cluster analysis provides us with a good way of understanding what kind of representations neural networks can encode. The hierarchical clustering generates a 'static symbolic description of a network's knowledge'—cf. Clark (1993). However, we should interpret this symbolic treatment as an external abstraction posited from the outside in an attempt to understand what the network is doing. Symbolic understanding is genuinely alien to the network *itself.* What gets activated, at each different time of rabbit-related processing, is a component of the cluster RABBIT*, not the cluster itself. In this way, we shouldn't see the network as actually using the context-independent concept RABBIT*. The network works exclusively at the level of the numerous context-dependent and distributed internal states of the hidden units. Uniting some of these states and putting them under the label RABBIT* should not drive us into thinking that the network actually employs the concept RABBIT*.

---

[49] That individual lexical concepts, identified as classical constituents, are causally efficacious is not necessarily common currency among sympathizers of the classical approach. Fodor, not surprisingly, still battles for the genuine causal efficacy of classical concepts—for his latest effort in this direction, see Fodor (1998a). The forthcoming discussion, thus, should be read as dealing exclusively with Fodor's particular approach. Other, less radical, classical (or classical-cum-connectionist) approaches, according to which the causality of mental states is not genuine but *derivative*, can be easily reconciled with my line of argument—see chapter 7 below.

In like vein, with respect to the simulation run by Elman reviewed in the first anti-Quinean rejoinder, we may say that the network *only* discovers 29 lexical items (the ones shown in figure 4.6). Symbolic knowledge of abstract NOUN and VERB categories is alien to the network. Positing those categories merely reflects the modeler's 'invasive' strategies to appraise the network's highly distributed representational resources. Unfortunately, the issue needs further elaboration, and a full appraisal of its implications must await until chapter 7, where I shall elaborate on the notion of causality in order to argue against the posits of folk psychology. Let me move now to the final stage of this chapter, where I shall try to wrap up my connectionist defence of the Quine's Inscrutability Thesis by looking at one interesting debate between the friends and foes of connectionism.

## 4.8 *Systematicity, Compositionality, and the Generality Constraint*

Fodor (1994, p. 295) says he is inclined to take LOT very seriously *if only* for lack of alternative candidates. As Fodor would put it, LOT is 'the only game in town'. Obviously this is not a bold assertion. Fodor has argued at length on the reasons why he believes connectionism cannot play the role LOT does as part of a representationalist theory of the mind (see references below). In a nutshell, Fodor's most powerful response to connectionism is that crucial aspects of cognition require a symbolic explanation. In particular, Fodor argues, connectionist models will not be able to explain the systematicity, productivity, compositionality, and inferential coherence of thought, unless they *implement* classical models, in which case LOT

wins. The issue is a thorny one, and a full appraisal would take us far afield.[50] In this section I shall focus exclusively on how to account for the systematicity and compositionality of thought whilst avoiding the charge of mere symbolic implementation. Addressing this issue is essential to reaffirm the credentials of connectionism as a genuine alternative to LOT *at the cognitive level*.[51] Moreover, reviewing the classical/connectionist debate will allow us to return to one of the two caveats left unanswered in chapter 2.

An origin of Fodor's misgivings towards connectionism as a genuine model of human cognition can be found in Chomsky's works in linguistics. Let me put the issue in perspective, and briefly sketch the traditional debate in linguistics on 'rules of grammar' as taken by Quine and Chomsky, respectively.

In *Methodological Reflections on Current Linguistic Theory* Quine posed a problem to linguists and philosophers alike: If we can be in possession of at least two extensionally equivalent systems of grammar for a given language *L*—i.e., two systems that can produce, by different routes, the same set of well-formed theorems of *L* in a recursive way—on what basis do linguists claim that one grammatical method with its own set of derivation-rules is the *correct* one? To make a long story short, Quine holds that we are not in a position to make such a choice. Correctness

---

[50] The reader may care to consult the *locus classici* of the classical/connectionist debate: Fodor and Pylyshyn (1988), Smolensky (1988), Fodor and McLaughlin (1990), and Smolensky (1991). See also Pinker and Prince (1988) for a powerful critique of connectionism which focuses, not on Fodorian LOT-like arguments, but rather on language acquisition in children.

[51] Note that Fodor wouldn't disagree with the connectionist that plausibly the brain is similar at the neural—architectural—level to connectionist networks.

is simply a matter of 'fitting' the behaviour of the speakers of *L*, and insofar as the different alternatives are extensionally equivalent the choice will remain unresolved.[52] By contrast, Chomsky (1957; 1968) maintains that being extensionally equivalent does not make two systems of grammar equally correct. Quine's behavioural constraint is too weak. There must be some further constraint upon a putative theory which explains how we are able to deliver an infinite number of well-formed theorems out of a finite cluster of axioms. This requirement drives us to Chomsky's notion of 'creativity of language' as accounted for by means of a *generative grammar*. Roughly, linguistic structures are generated by applying recursive sets of generative grammatical rules to other linguistic components. Application of these generative rules, Chomsky argues, is capital for linguistic production and comprehension. Granting this framework, Chomsky can argue *contra* Quine in favour of a *particular* generative grammar.[53] Apart from achieving the correct output—i.e., delivering the correct set of well-formed sentences—we must also contemplate the *way* we arrive at those sentences. Thus, Chomsky contends, there is only one correct system of rules. Namely, the one that competent speakers have somehow implicitly in mind, and which informs their judgment as to which utterances are grammatical, and which are not. Chomsky aims to answer

---

[52] What counts as *fitting the evidence* for Quine would require some further qualification. Without worrying about the fine-grained detail, we may venture to say that to fit the evidence in the theory of grammar is to recognize certain strings of noises as grammatical—see Quine (1972); and chapter 1, above.

[53] This is a simplification of Chomsky's argument. For the *minutiae* the reader may care to consult Chomsky (1968).

Quine's original skepticism about realism towards systems of rules of grammar. A correct generative grammar is the one that tacitly drives the speakers via a canonical route from the axioms of the theory of grammar to its theorems. Bearing in mind this setting, we may now extend these ideas to the theory of mental representation.

Chomsky's work in linguistics, together with the impressive results of digital computers, supported Fodor's view of cognition as manipulation of symbols according to rules explicitly realised in the brain—see 4.3 above. To remind the reader, LOT crucially has a combinatorial syntax and semantics. Molecular representations are formed out of smaller constituents. Sentence crunching requires the sentences to have constituent structures, where the rules of crunching are syntactic. That is, the rules for combining or decomposing mental representations can be applied without regard to the semantic character of the symbols involved. What we have is thus that mental representations are to be seen as linguiform complex representations whose semantic properties are directly determined via the semantic properties of their constituents; the simplest constituents, then, form a stock of *context-independent* items. The combination and recombination of these content-bearing representational units allow thinkers to entertain and grasp novel thoughts.

According to Fodor's Syntactic Image, the radical linguist of Quine's parable (see 1.3) would see natives as tacitly interpreting linguistic rules which can be explicitly formulated in a theory of grammar. Since distributed neural networks encode information about sentences without employing LOT's stock of context-independent items, the sympathizer of LOT will argue that connectionism cannot

explain the systematicity and compositionality of thought. Unfortunately for the sympathizer of Fodor, were connectionism to earn its keep empirically (see 4.5 above), the linguiform metaphor of mental representation, and in particular the requirement of explicit linguistic-like rules for the combination and recombination of context-free constituents, seems to be at least superfluous. As I concluded in section 4.6, lexical inclusion in natural languages does not imply conceptual inclusion when we move to the theory of mental representation. Under the connectionist approach, the finite set of basic clauses of a translational manual do *not* give a basic repertoire of concepts from which all other concepts are construed. The connectionist basic-to-phrasal direction of conceptual formation is orthogonal to the requisites imposed by LOT's classical constituency. The basic clauses are lexically basic, but have no privileged conceptual status. What the connectionist then needs is a way to account for the abstract structure that seems to underlie the systematicity and compositionality found in thought processes, but without making use of LOT's classical form of constituency. That is the target of the remainder of this section. Recent research due to Elman (1998) seems to imply that the Chomskian approach to linguistics, and Fodor's subsequent extension to thought, are on the wrong track.

Elman (1998) trained a network to answer a criticism, along the lines of Fodor and Pylyshyn's (1988), put forward by Hadley (1992), and Marcus (1998). The challenge posed by these authors is to explain how connectionist networks can account for *strong systematicity*. Strong systematicity refers to a sort of generalization in which the network or cognitive agent must generalize to

previously unencountered grammatical roles. For example, given a network trained on sentences in which, say, the noun 'boy' *only* plays the role subject, the question is whether the network can deal with novel sentences when 'boy' plays the role 'object'. Elman's simulation, as we shall see next, shows that neural networks can account for this form of strong systematicity. These promising results will shed some new light on the classical/connectionist debate. Let me briefly review the key aspects of Elman's simulation.

Elman (1998) trained a simple recurrent network on a prediction task along the lines of the simulations reported earlier—Elman (1990; 1992), sections 4.6, and 4.7 above. The task is to predict successive words in sequences of sentences. The key point, this time, is that words have different probability of occurrence in the grammar (see Elman, 1998, for the details). To address the challenge of accounting for strong systematicity, Elman focused on the noun 'boy'. Given that 'boy' never appeared in direct object position for any verb in the training data set, can the network predict 'boy' as a direct object after presentation of the verb 'talk-to'? Generally speaking, the challenge is whether the network can successfully generalize to previously unencountered grammatical roles. As Elman's research shows, the network does manage to account for this form of systematicity. In particular, the network predicts 'boy' in the context 'the girl talks to ...', even though the network never saw 'boy' in any object position during training.

Elman points out the reason for this exciting result: It is vital for the network to predict 'boy' in the context 'the girl talks to ...' that 'boy' has already been fed to the network during training in other contexts *together with* other human words—

e.g., girl, man, woman. So, for example, as Elman argues, in the toy-language employed for this simulation only human words appear in subject position with verbs such as 'eat', 'give', or 'transfer'. On the other hand, neither 'boy' nor other human words appear in object position with verbs such as 'terrify' or 'chase'. In short, even though the network never sees 'boy' in an object position, it is trained on roles that 'boy' *shares with* other human words (more than it does with other types of words). These 'behaviour-based similarity'—see Elman, 1998— between 'boy' and other human words is what allows Elman's network to generalize to previously unencountered syntactic roles.

What is noteworthy about Elman's (1998) network is that it succeeds in its overall target—i.e., accounting for strong systematicity—without implementing classical rules of the sort Fodor sees as necessary. It is important to emphasize the reason for this. As I mentioned in section 4.5 above, all the knowledge the network acquires is *superimposed* on the same hardware (see also 4.9 below). This allows us to understand the network's capability of generalizing to unencountered syntactic positions. When the network is being input a new activation pattern for 'boy', the output for other non-related types of words (e.g., 'dragon') remains largely unchanged, and *vice versa*. The reason is that the weight changes are distributed over the entire set of connections. Therefore, since the network's representation of 'dragon' is significantly different form the one of 'boy', new information about

'dragon' will have minor repercusions on the representational storage of 'boy'.[54] On the other hand, given that the activation patterns for, say, 'girl', 'man', or 'woman' are very similar to the one encoding for 'boy', there will be a high correlation between weight changes and activation patterns for tokens of these word-types. In this way, any new piece of information about 'girl', 'man', or 'woman' is automatically generalized to 'boy', to the degree that the representations for 'girl', 'man', and 'woman' are similar to the one for 'boy'. Bearing this in mind, the behaviour of the network, nonetheless, could still 'economically' be described in terms of classical rules. However, the network is alien to the spirit of the *Syntactic Image* (see 4.3). It employs neither grammatical classical constituents, nor is the processing sensitive to the syntax of such constituents. The behaviour of the network remains rooted in representations which are context-dependent, driven by the inherently dynamical character of the interconnection of many simple units as explained above, and in sections 4.4-4.6.

Before moving on to the issue of compositionality, let me expand briefly on how Elman's results should be interpreted as far as the debate on *rules* goes. This will be crucial to appraise fully the classical/connectionist debate. It would not be accurate to claim that we can account for the systematicity of thought without following rules. Neural networks do follow rules. However, the reader should not

---

[54] As a matter of fact, the changes required to encode new information about 'dragon' will have a random effect on 'boy'. The result is that potential representational effects will cancel out when averaging over many trials.

infer any similarity between classical and connectionist rules.[55] Given an input domain in a training space, and a range of possible outputs, the network's task is 'simply' to find a successful approximation of the input/output function. But an input/output function is nothing but a rule—see Elman (1990). What is really at stake then is whether the rules employed by connectionist models are equivalent to the sort of explicit rules that classical models make use of. The learning rules employed by neural networks concern exclusively how the weights will change as a result of an incoming flow of activation. The mathematical description of such changes bears no resemblance to the explicit rules being stored and retrieved by classical machines. Connectionist networks behave in a rule-like manner exclusively in virtue of the readjustment of connection weights according to learning algorithms—see 4.4 above. The illusion of classical-rule-*governed* behavior is a result of the *non-linear* character (4.4) of the activation function performed on the hidden units. The non-linear response of connectionist networks means that under certain circumstances hidden units react in an all-or-nothing fashion, and under certain other circumstances, they react continuously (see figure 4.1 above). This kind of non-linear response is what permits units to behave in a categorical, rule-like, manner.[56] In short, the key point is that even though connectionist processing is functionally equivalent to classical processing, the representations and rules that connectionist networks make use of are highly

---

[55] The key issue is not whether connectionist networks employ rules or not. This, at times, seems to be the issue Fodor and Pylyshyn (1988) are concerned about.

[56] For an exhaustive treatment of this topic see Elman *et al.*, 1996, chapter 4.

distributed and context-dependent.[57]

As we saw earlier, Chomsky's generative grammar posits sets of rules in order to account for linguistic productivity. The brain, he assumes, follows those rules. If the above connectionist picture is correct, the brain does follow rules, although rules rooted at the weights-and-units level of processing, and completely orthogonal to classical constraints. A further constraint that Fodor reads off Chomsky's approach to linguistics is the requirement of semantic compositionality—indispensible if a semantic theory is going to be able to deliver an infinite number of theorems out of a finite set of axioms. As I argued earlier, assuming connectionism, the finite set of basic clauses of a translation manual, or a theory of semantics, does not furnish us with a basic repertoire of concepts from which all other concepts are construed. It seems then that Fodor's demand for compositionality is orthogonal to the connectionist semantic enterprise. Again, as in the discussion on rules, further clarification is needed.

Connectionism does not drop the need to account for compositionality, but rather it simply drops its classical reading. Following van Gelder (1990), I shall distinguish between two forms of compositionality: *concatenative* compositionality, and *functional* compositionality. Concatenative compositionality requires the preservation of tokens in order to build up the increasingly complex structures of our mental representations. Complex representations can only be tokened by

---

[57] The reader may care to consult Rumelhart and McClelland's (1986) neurosimulation on English past tense acquisition for a neat illustration of how the English past tense formation can be mastered without recurring to explicit rules to discern between regular and irregular verbs.

tokening their constituents. Concatenative compositionality—in van Gelder's terminology—amounts to the classical form of compositionality championed by Fodor. Connectionist models are incompatible with this kind of compositionality. According to Fodor, complex thoughts arise from the combination and recombination of *classical*—content-free—constituents (see 4.3 above). To remind the reader, Fodor's commitment to classical constituency amounts to the claim that:

> for a pair of expression types $E_1$, $E_2$, the first is a *Classical* constituent of the second *only if* the first is tokened whenever the second is tokened. (Fodor and McLaughlin, 1990, p. 186)

The reader can see why connectionism is incompatible with concatenative compositionality. As we saw in section 4.6, connectionist networks lack a classical form of constituency. The question then is: How can connectionism account for the compositional character of thought, given that connectionist representations employ context-dependent constituents. The answer comes by the hand of van Gelder's second type of compositionality—*functional* compositionality.[58] Functional compositionality does not require the preservation of constituents. According to van Gelder,

> [functional] compositionality is obtained when there are general, effective, and reliable processes for (a) producing an expression given its constituents, and (b) decomposing the expression back into those constituents. (Van

[58] For the reader unfamiliar with the literature, van Gelder's functional compositionality amounts to what Elman *et al.* (1996) dub 'interactive' compositionality.

Gelder, 1990, p. 361)

The combination lock metaphor (van Gelder, 1990) neatly illustrates the underlying mechanisms of functional compositionality. Think of the numbers in the combination of a lock as playing the role words do in languages. Number's causal properties are context-dependent. That is, a correct *sequence* of numbers gets the lock opened. We may then say that the final state—i.e., lock being open—is compositional since it is dependent on a specific sequence of numbers. Numbers, thus, although not preserved physically, are somehow 'functionally present'. In like vein, we may say that although individual concepts cannot be preserved across context in connectionist networks, they are still in the system.

Connectionist constituency, as found in recurrent networks, can account for the processing of natural languages, and the representation of complex hierarchical grammatical structures. The neurosimulations reported in this chapter have the capacity to account for the temporal dimension involved in processing sequential inputs, and can do so while meeting van Gelder's (a) and (b) *desideratum* in the above quote. In this way, I conclude, connectionism with its non-classical form of constituency, and non-linear processing, can account for the systematicity and compositionality of thought. Given these results, I shall argue next, Evans' generality constraint becomes innocuous against the Quinean.

As I mentioned in chapter 2 (section 2.3), Evans is confident that considerations regarding the productivity and systematicity of language and thought will definitely discredit perverse semantic theorizing à la Quine. To remind the reader, the pivotal factor in Evans' argument is the existence of the generality

constraint; a constraint to the effect that:

> if a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception. We thus see the thought that *a* is *F* as lying at the intersection of two series of thoughts: on the one hand, the series of thoughts that *a* is *F*, that *b* is *F*, that *c* is *F*, ..., and, on the other hand, the series of thoughts that *a* is *F*, that *s* is *G*, that *a* is *H*, ... (Evans, 1982, p. 104)

Evans' generality constraint stresses the need to posit *single* causally efficacious states in order to explain the regularities manifested in linguistic behaviour. In short, the generality constraint calls for a single inner state which gets activated whenever a cognitive episode involving a given concept occurs. This view is congenial with Fodor's classical approach, and therefore threatens Quine's Inscrutability Thesis. In chapter 2 I postponed addressing the issue of whether the generality constraint can help the anti-Quinean to narrow down the range of empirically adequate semantic theories. Given Evans' full-blooded approach to semantics—i.e., granting that linguistic comprehension is to be accounted for at the *neurophysiological* level (see chapter 2, fn. 6)—the issue remained an open question. I believe, however, that the preceding discussion on concept naturalization (sections 4.3-4.7), and the above discussion, furnishes us with the perfect tool kit to tackle Evans' 'generality constraint' challenge.

Evans' generality constraint exploits the extraction of *regularities* typical of symbolic models of cognition. However, as we saw earlier, connectionism does not

need to posit *single* causally efficacious states to account for linguistic regularities. Connectionist constituents are differentiated by virtue of playing different causal roles, ever rooted at the level of idiosyncratic processing. This, nonetheless, does not suppose a shortcoming for the connectionist. If the above considerations are on the right track, and we can account for the productivity of language and thought via a functional form of compositionality, then the worry arising from Evans' generality constraint is harmless. In particular, we shall be able to explain any set of inferences without having to posit an internal factor which is *common* to all inferential transitions. Complexity, through the connectionist lens, emerges from the non-linear properties of simple dynamical systems. Evans' call for a common piece of concept mastery becomes, I claim, an artifact created by endorsing a classical model of cognition.

## 4.9 *Conclusion*

Before closing this chapter let me address a minor source of worries drawn to my attention by U.T. Place. Place is keen on favouring connectionism as a genuine alternative to the Fodorian syntactic image. However, Place (e-mail communication) disagrees with me with regard to the implications of endorsing a connectionism model of cognition for Quine's Inscrutability Thesis. Human language, Place contends, is superimposed on an elaborate system of pre-linguistic concepts, part innate and part learned (but not requiring a language of thought hypothesis). The function of this system is to enable organisms to recognize things

of kind (and individuals for that matter) in the sense of pre-selecting a range of behavioural strategies appropriate to encounters with things of the kind in question. In such a system of concepts, universals like that to which we assign the word 'rabbit', which pick out a common biological species likely to be significant as prey to any larger predator, are likely to take precedence over a universal such as, for example, Quine's 'undetached rabbit part'—see chapter 1. The reason, Place argues, is simply that having a set of behavioural strategies appropriate for dealing with the former is going much more useful to an average predator than is the latter.

I don't think that Place's considerations can endanger our perverse semantic theory, $PT_4$ (see chapter 3). Indeed, the complaint we can see emerging from Place's comment did not escape Quine's notice. Plausibly, there is a positive constraint between 'behavioural strategies' and 'significant survival value'. As Quine notes:

> Man is a body-minded animal, among body-minded animals. Man and other animals are body-minded by natural selection; for body-mindedness has evident survival value in town and jungle. (Quine, 1973, p. 54)

Hence, if a predator targets whole enduring rabbits it will surely have more chances to survive and pass those hunting skills down to future generations, than if it chases small undetached rabbit parts (such as a rabbit's claw), ignoring the rest of the rabbit. However, a predator being *body*-minded should not be confused with being *rabbit*-minded, for instance. Quine's notion of 'body-mindedness' is not that restrictive—as a matter of fact, that's the very point at stake. As Quine points out following the above passage:

> When the time comes for the precision of physical science, the notion of body can give way to the more inclusive, more recondite, and more precise notion of *physical object.* Any arbitrary congeries of particle-stages, however spatiotemporally gerrymandered or disperse, can count as a physical object.[59] (Quine, 1973, p. 54)

In this way, by favouring a loose notion of 'body-mindedness', the only problem left for the perverse semanticist is to reconcile being 99% body-minded with having survival value. But this should not be cause of concern for the Quinean. Survival value is partly determined by the afferent/efferent connections possessed by organisms. Afferent/efferent connections appropriately linked to certain environmental features lead to better chances of survival. Environmental features can be naturalistically/evolutionarily anchored by looking at the production of certain efferents in *differential* response to afferents. At this point, I claim, what defines a given stimuli cannot help distinguish between, say, rabbits and 99%-urp. The causal chain leading to a given afferent/efferent pattern stretches back, through several stages, to the image of a rabbit in the retina, to certain patterns of light rays produced on the rabbit's surface, and finally, to the rabbit itself. Also, we are designed by evolution to maintain a degree of constancy between the several representational stages that occur between afferent and efferent response, and the perceived objects themselves. However, evolutionary arguments cannot help to make the referential indeterminacy urged by the Quinean dissipate. Note that, were we to favour the '99%-urp' referential scheme, we would discover that the chances

for the predator to survive are as high when it chases a 99%-urp as they are when the whole rabbit is being chased. *In terms of survival value*, there is no real significance between rabbits and 99%-urp.[60]

The purpose of this chapter has been to show that considerations urged by Wright (1997) regarding complexity in the psychological theory that accompanies semantic theorizing are unable to discredit perverse semantic theorizing. In a worst-case scenario for the Quinean, a standard interpretation of Native might be favoured. However, I argued, the price to pay would be the endorsement of a LOT hypothesis. Research in neurobiology and cognitive science *appears* to discredit the LOT hypothesis, and favour a connectionist model of cognition. *If* connectionism is correct *then*, I claimed, Wright's 'psychological simplicity' criterion is unmotivated, favouring neither a standard nor a perverse interpretation of Native. Therefore, my connectionist defence of Quine's Inscrutability Thesis is dependent on whether future empirical research confirms or disconfirms the Fodorian LOT hypothesis. The issue, I admit, is still an open empirical question. At this point, thus, the best way to frame the results of this chapter is conditionally (see chapter 7 below).

Granting for argument's sake that connectionism is the correct model of cognition, the Quinean has more reasons to celebrate. The Inscrutability Thesis

---

[59] For an expansion on this point, see Quine (1973), §§ 23; 34. See also chapter 1 above.

[60] Place's point may nonetheless hold when the part being chased is for example a 1%-urp. However, I would need to see in more detail an argument along Place's lines against such perverse option before submitting it to critical scrutiny.

would not be the only Quinean thesis to obtain neuroscientific support. Connectionism appears to vindicate Quine's holistic approach to semantic content as well. Quine's semantic holism, in a nutshell, maintains that the relevant unit of meaning is not the word or the sentence, but rather the language—or the organism's cognitive theory—as a whole. Semantic holism is vindicated because of the *superpositional* character of connectionist representations. The basic idea is that fully distributed neural networks exploit superpositional storage techniques.[61] As I mentioned earlier (section 4.5) a *single set of weights* allows neural networks to constantly generate the right activation patterns in the face of the activation from new input patterns. Representations are said to be fully superposed if the resources the network employs to represent one item are the same as those required to represent a different item:

> Thus, if a network learns to represent item 1 by developing a particular pattern of weights, it will be said to have superposed its representations of items 1 and 2 if it then goes on to encode the information about item 2 by amending the set of original weightings in a way which preserves the *functionality* (some desired input output pattern) required to represent item 1 while simultaneously exhibiting the functionality required to represent item 2. A simple case would be an autoassociative network which reproduced its input at the output layer after channeling it through some intervening bottleneck (such as a small hidden-unit layer). Such a net might need to find a simple set of weights which do multiply duty, enabling the net to reproduce any one of a whole set of inputs at the output layer. If all the weights turned out to be playing a role in each such transition, the representation of the

[61] For the reader interested in expanding on superposition the *locus classicus* is Van Gelder (1991).

various items would be said to be *fully* superposed. (Clark, 1993, p. 17)[62]

Superpositional processing of all the existing informational states of a network fit perfectly with the thesis of semantic holism. The content of a concept, as described in connectionist terms, would be determined by the superposition of all the available representational resources, which are a function of the whole range of input/output patterns that the network has been trained on. It seems then that Quine's overall behaviouristic position fits like hand in glove with the neurophysiological level of explanation provided by connectionist recurrent networks.

I conclude then that the Quinean can go with modern scientific fashion and make use of the '99%-urp' referential scheme. Unfortunately, the anti-Quinean has one more rejoinder up her sleeve. This other criticism exploits recent experimental research which highlights the existence of an objective criterion of conceptual similarity in connectionist terms. The anti-Quinean then hopes that a connectionist sympathizer of Wright may still manage to press on his 'psychological simplicity' argument by submitting standard and perverse concepts to the test of 'conceptual similarity'. The anti-Quinean will argue that standard concepts are more similar to their Native counterparts than perverse ones are. In my opinion this criticism is far more serious than those previously addressed in this chapter. In the next chapter I shall expand on this criticism and offer a solution which, I hope, succeeds in

---

[62] See Van Gelder, 1991, p. 43, for a more technical definition of fully superpositional representation.

retaining the empirical adequacy of the perverse semantic theory offered in chapter 3 above.

# 5

# *STATE SPACE SEMANTICS AND CONCEPTUAL SIMILARITY*

## 5.1 *Introduction*

Jerry Fodor and Ernest Lepore (1992; 1996) have launched a powerful attack against Paul Churchland's connectionist theory of semantics—aka *State Space Semantics* (see chapter 4, section 4.5). In one part of their attack, Fodor and Lepore argue that the architectural and functional idiosyncrasies of connectionist networks preclude us from articulating a notion of conceptual similarity applicable to State Space Semantics. Aarre Laakso and Gary Cottrell (1998; 2000) have recently run a number of simulations on simple feedforward networks, and applied a mathematical technique for measuring conceptual similarity in the representational spaces of those networks. Laakso and Cottrell contend that their results decisively refute

Fodor and Lepore's criticisms. Paul Churchland (1998) goes further. He uses Laakso and Cottrell's neurosimulations to argue that connectionism does furnish us with all we need to construct a robust theory of semantics and a robust theory of translation. Although the Fodor-Lepore/Churchland debate concerns exclusively the metaphysical status of State Space Semantics (see 5.2 below), Churchland (personal communication) believes that the outcome of the debate—were connectionist semantics à la Churchland[1] to earn its keep—may have a negative bearing upon the connectionist defence of Quine's Inscrutability Thesis put forward in the previous chapter. In particular, Churchland contends that a connectionist sympathiser of Wright may be able to exploit Laakso and Cottrell's neurocomputational results in order to vindicate a version of Wright's 'psychological simplicity' argument, thus putting in jeopardy any Quinean perverse semantic theory (see chapter 4, section 4.2, above). In this chapter I shall argue that whereas Laakso and Cottrell's neurocomputational results may provide us with a rebuttal of Fodor and Lepore's argument, Churchland's conclusion is far too optimistic. In particular, I shall try to show that connectionist modeling does not provide any objective criterion for achieving a one-to-*one* accurate translational mapping across networks, as the foe of Quine requires.

---

[1] Other connectionist approaches to the theory of semantics, and the theory of mental representation, seem to avoid the criticisms put forward by Fodor and Lepore that I'll review in this chapter. For present purposes I shall concentrate in Churchland's proposal, ignoring other connectionist, maybe more fruitful, semantic proposals. For a general appraisal of the landscape, and how connectionist semantics can be given its best shot, see Tiffany (1999).

Before getting started, let me briefly outline the programme of this chapter. In section 5.2 I shall briefly review State Space Semantics, and what the problem for the theory is, all according to Fodor and Lepore. In section 5.3 I shall introduce a mathematical technique for measuring conceptual similarity across networks that Laakso and Cottrell have recently offered in order to address Fodor and Lepore's challenge. In section 5.4 I shall show how Churchland makes use of Laakso and Cottrell's results to argue that connectionism can furnish us with all we need to construct a robust theory of semantics, and a robust theory of translation—robustness that may potentially be exploited by a connectionist foe of Quine to argue against the Inscrutability Thesis. In section 5.5 I shall argue that Churchland's conclusion is far too optimistic. In particular, I shall try to show that the notion of conceptual similarity available to the connectionist leaves room for a "connectionist Quinean" to kick in with a one-to-*many* translational mapping across networks. In section 5.6 I shall highlight a potential problem for Laakso and Cottrell's rebuttal of Fodor and Lepore's criticism, and Churchland's subsequent defence of State Space Semantics, that has been completely ignored in the connectionist literature. Conclusions and suggested directions for future research will follow in section 5.7.

## 5.2 *State Space Semantics: The Problem*

As we saw in chapter 4, Churchland (1986) has proposed a new, connectionist-inspired, approach to the theory of mental representation known as State Space Semantics. Briefly, the basic idea behind Churchland's proposal was that

> [the] brain represents various aspects of reality by a position in a suitable state space, and the brain performs computations on such representations by means of general coordinate transformations from one state space to another. (Churchland, 1986, p. 280)

Churchland invites us to view *concepts* as points in a partial state space of a dynamical system. These points correspond to the tips of the vectors determined by the levels of activation of the different units in hidden layers. The semantic characteristics of a concept can then be seen as a function of the *place* that that concept—i.e., point—occupies in a geometrically characterized hyperspace. In this way, Churchland proposes, we may talk of semantic similarity between concepts in terms of the proximity of their respective *absolute* positions in state space, as identified in relation to a number of semantically relevant dimensions.[2]

Jerry Fodor and Ernest Lepore (1992; 1996) have recently launched a powerful attack against Churchland's proposal. One of their objections can be

---

[2] The reader not familiar with the basic tenets of connectionist theory is urged to visit chapter 4, sections 4.4-4.6, above.

summarized as follows:[3] State Space Semantics understands conceptual similarity across networks as similarity in the activation patterns across those dimensions that specify the networks' representational spaces (see 4.4 above).[4] However, under this connectionist framework, it seems that two individuals—i.e., networks—cannot possibly entertain the same concept. And the reason for this is that processing in connectionist networks is highly idiosyncratic. Differences, for instance, in the *encoding* of the input data, in the *architecture* of the model, and in the *dimensionality* in hidden space, strongly constrain *how* a network proceeds in order to achieve successful performance. Learning, in short, is highly sensitive to the idiosyncrasies of neuromodeling. These considerations have driven Fodor and Lepore to argue against State Space Semantics as a putative theory of mental representation. Idiosyncrasies in encoding, architecture, or hidden dimensionality make it impossible to talk of similarity of patterns of activation across networks. It then seems to follow straightforwardly, Fodor and Lepore argue, that we cannot talk either of similarity of positions in state space across networks. It is important

---

[3] What follows is a simplification of one part of Fodor and Lepore's argument. Although for our present purposes it will suffice. For an appraisal of Fodor and Lepore's overall argument against State Space Semantics, the reader may care to consult the exchanges between Fodor and Lepore, and Churchland in McCauley (1996), and Fodor and Lepore (forthcoming). For a defence of State Space Semantics, see Tiffany (1999). For a rebuttal of Churchland's general strategy to bypass Fodor and Lepore's criticism see Calvo Garzón (in preparation b).

[4] Just a word on notation. In what follows, I shall employ the terms 'activation pattern', 'vector, and 'point' interchangeably as referring to one and the same thing. Namely, to the unit of representation in connectionist semantics.

however to emphasize the root of their distrust. Fodor and Lepore's claim is *not* that connectionism cannot define what it is for two individuals to entertain similar concepts. Their claim is *not* that connectionism lacks a measure *to judge* whether different individuals represent a given input in the same conceptual way. Fodor and Lepore write:

> If the paths to a node are collectively constitutive of the identity of the node, [...] then only identical networks can token nodes of the same type. Identity of networks is thus a sufficient condition for identity of content, but this sufficient condition isn't robust; it will never be satisfied in practice.[5] (Fodor and Lepore, 1996, pp. 146-7)

As this quote illustrates, Fodor and Lepore are not denying the logical point that we can have a connectionist measure of conceptual similarity—see section 5.3 below. Their point is rather ontological—viz., that the conditions for conceptual similarity set out by State Space Semantics will never allow two individuals to share a given concept (given that human brains have different numbers of neurons, which are differently connected to each other, and which exhibit different patterns of causal connectivity).

Churchland does not seem to be moved by Fodor and Lepore's criticism:

---

[5] At this point, Fodor and Lepore are actually targeting the classical Quinean "web" picture of theories/languages/belief systems, in order to argue that it cannot provide a robust account of conceptual *identity*. However, the argument applies equally to State Space Semantics, and its incapability to furnish us with a robust notion of conceptual *similarity*—see Fodor and Lepore, 1996, pp. 146-ff.

> The short answer to [Fodor and Lepore's] critique is that content is not, in general, assigned in the manner described. A point in activation space acquires a specific semantic content not as a function of its position relative to the constituting *axes* of that space, but rather as a function of (1) its spatial position relative to all of the *other contentful points* within that space; and (2) its causal relations to stable and objective *macrofeatures of the external environment.* (Churchland, 1998, p. 8)

Churchland hopes to bypass Fodor and Lepore's attack by equipping State Space Semantics with a *non-absolute* measure of conceptual similarity. As we saw earlier, patterns of activation get their content as a function of the content of the dimensions that define the representational space in question. Conceptual similarity across networks was then defined in terms of the similarity of the *absolute* positions within each state space. By contrast, Churchland now puts the emphasis on the similarity of the *relative* positions of different activation patterns. We may then define conceptual similarity across networks in terms of the position of a given pattern of activation in relation to other patterns in the same representational space. In this way, we may say that two networks share the same conceptual repertoire if the *set of relations among the activation patterns* in the first network is isomorphic—see section 5.4 below—to the set of relations obtained in the second network.

Churchland's new account shows some promise in the fact that a non-absolute definition of similarity relaxes the demands on State Space Semantics. Note that now we can ignore the different dimensionality, as well as the particular microcontent of each dimension of each state space. All we need then—or so it

appears to Churchland—is to establish a set of necessary and sufficient conditions for a *relative definition* of conceptual similarity. To achieve these goal, Churchland turns to some empirical research carried out by Laakso and Cottrell. That research is the subject matter of the following section.

## 5.3 *A Connectionist Measure of Conceptual Similarity*

Laakso and Cottrell (1998; 2000) have recently taken up Fodor and Lepore's challenge (see 5.2 above). According to Laakso and Cottrell, we do have a criterion for judging conceptual similarities across different connectionist networks. Namely, by measuring distances among points within the hidden space of a given network, and correlating those measures with the measures obtained within the hidden space of a distinct network. They illustrate their strategy with a simple case—see Laakso and Cottrell (2000). Take two networks—network #1 and network #2—with one and two hidden units, respectively. Both networks learn to represent three unspecified things, say A, B, and C. Network #1 represents A, B, and C with the following vectors:

A = <0>, B = <50>, and C = <100>.

On the other hand, network #2 represents the same three things with the following vectors:

A = <0, 0>, B = <30, 30>, and C = <80, 0>.

We can then form the following matrices (see fig. 5.1 below) by considering the distances between all the representations within network #1, and also comparing the distances between all the representations in #2. Now, by computing these distances, we can employ a mathematical measure of *similarity* with which to compare the representations of networks #1 and #2. Since both matrices are symmetric we can extract the respective vectors and compare them.

| Distances Between Representations | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1-Unit Network | | | | 2-Unit Network | | | |
| | A | B | C | | A | B | C |
| A | 0 | 50 | 100 | A | 0 | 42 | 80 |
| B | 50 | 0 | 50 | B | 42 | 0 | 58 |
| C | 100 | 50 | 0 | C | 80 | 58 | 0 |

[Fig. 5.1]: Symmetric matrices obtained by taking Euclidean distances between all the representations in each network. (From Laakso and Cottrell, 2000)

In our example, the two vectors are:

<50, 100, 50>, and
<42, 80, 58>

which, having the same dimensions, can be easily compared. The idea, in short, is that points in different hidden spaces stand for the same, or similar, things in case there is a high *correlation* between the distances among the sets of points—i.e.,

concepts—in the respective networks. With this mathematical measure, Laakso and Cottrell argue, we need not worry about Fodor and Lepore's argument. Different dimensionality, architecture or encoding bring no trouble, insofar as correlated distances between points in the respective spaces are preserved.

Laakso and Cottrell tested this strategy in two different experiments. In the first experiment, they trained several three-layer feedforward nets, all containing three hidden units, on a colour-categorization task. The networks were trained using four different input encodings. The outputs were: "red", "yellow", "green", "blue", and "purple". After obtaining the activation patterns at the hidden layer for each different input pattern, Laakso and Cottrell computed the Euclidean distances between each different pair of activation patterns in hidden space for a given net. Finally, they compared the activation patterns in the two nets by computing the correlations among the hidden activation patterns obtained in each net. Laakso and Cottrell reported that the representations obtained for every input presented were highly correlated across networks.

Though an important result as it is—think of the various input encodings as corresponding to different species' sensory modalities—all the networks contained the same number of hidden units, and thus did not fully address Fodor and Lepore's challenge. Laakso and Cottrell then ran a second experiment, again on a colour-categorization task, but this time employing networks with different internal dimensionality, as well as different input codings. The networks employed had between 1 and 10 hidden units. Once the networks mastered the categorization task, the mathematical measurements were computed as above, and as in the previous

experiment, the correlations obtained were very high, independently of the number of hidden units employed by the networks.[6] From these results, Laakso and Cottrell conclude:

> Our measure is a robust criterion of content similarity, of just the sort that Fodor and Lepore demanded in their critique of Churchland. It can be used to measure similarity of internal representations regardless of how inputs are encoded, and regardless of number of hidden units. Furthermore, we have used our measure of state-space similarity to demonstrate empirically that different individuals, even individuals with different "sensory organs" and different numbers of neurons, may represent the world in similar ways. (Laakso and Cottrell, 1998, pp. 595-6)

Laakso and Cottrell's results get connectionist semantics off the ground, and seem to shed new light on the Fodor-Lepore/Churchland debate over the fate of State Space Semantics.[7] The question I would like to pursue next is to what extent

---

[6] For the details of both experiments, see Laakso and Cottrell (1998).

[7] Just a word of caution. To keep the record straight, Fodor and Lepore's point is not an *epistemic* one. What *can or cannot be judged, or measured* is not what's at stake—see section 5.2 above. Both Laakso and Cottrell (1998), and Churchland (1998) seem, at times, to be taking Fodor and Lepore to be presenting an epistemic challenge. So, for example, commenting on Laakso and Cottrell's strategy, Churchland writes: "The *truly important point is that we can tell* whether or not [various networks settle on the same cognitive configuration in response to their shared problems]. We can say what their internal cognitive similarity consists in, and we can give an objective numerical measure of that similarity" (Churchland, 1998, p. 24; my emphasis). In response to a previous version of this chapter—see Calvo Garzón (2000b)—an anonymous referee for *Philosophical Psychology* urges that Churchland's epistemic reading may be evading the real issue prompted by

Churchland can make use of Laakso and Cottrell's results to reaffirm the credentials of State Space Semantics as a *robust* theory of mental representation. In the remainder of this chapter I shall elaborate on this issue in order to argue that the metaphysical status of State Space Semantics may be worse than Churchland would be willing to admit. As a result, I shall contend, a connectionist foe of Quine won't be able to make use of Laakso and Cottrell's neurosimulations to reinforce Wright's 'psychological simplicity' argument.

## 5.4 *Similarity of Prototypical Trajectories: A Solution?*

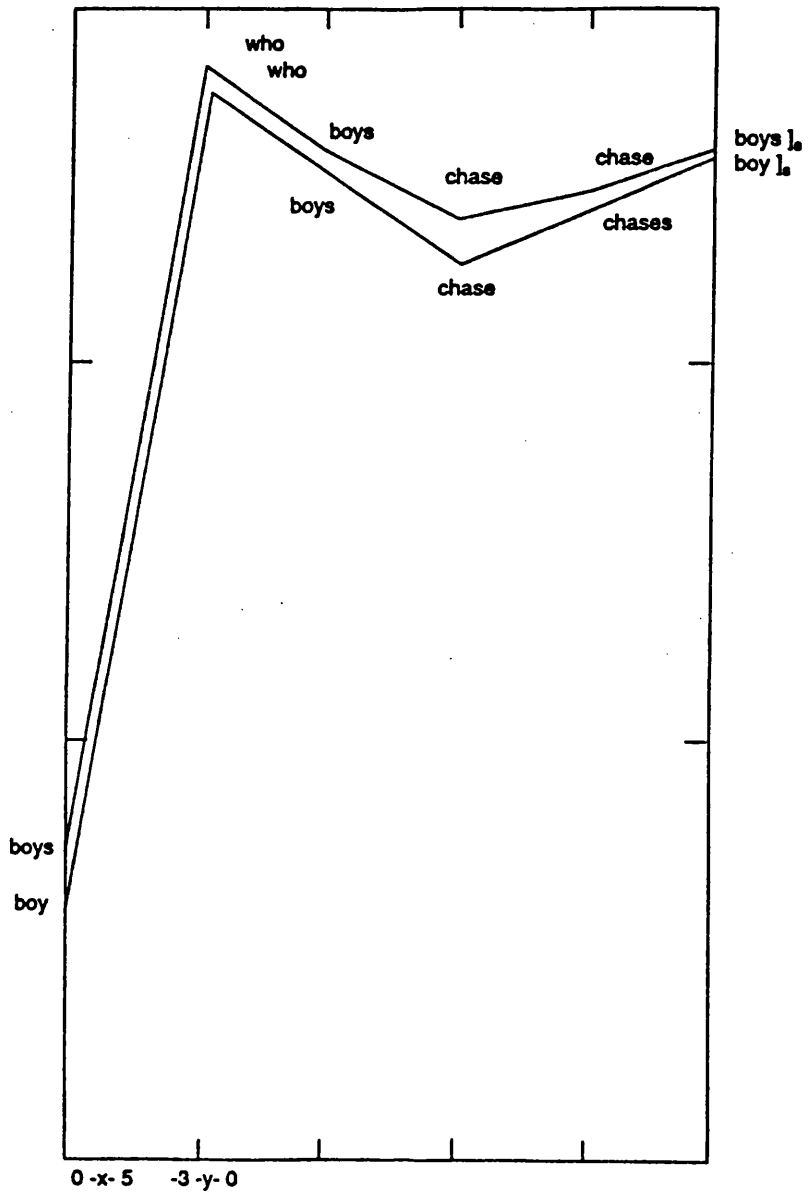
Laakso and Cottrell conducted their simulations with simple feedforward nets on a colour-categorization learning task. The output was a single word—either "red", or "yellow", etc. However, if we are to account for the whole range of *human* cognitive capacities, we need to expand Laakso and Cottrell's results, at least, to simple recurrent networks of the kind employed to process sentences belonging to a small portion of a natural language. As we saw in chapter 4 (section 4.6), Jeff Elman (1992) designed a simple recurrent network which exhibited appropriate sensitivity to the syntactical dependencies found in sentences. To remind the reader, a simple recurrent network, thanks to the employment of feedbackward pathways, can deploy some sort of short-term memory, that allows the network to process

---

Fodor and Lepore. Namely, to find a *robust* notion of conceptual similarity. For present purposes, however, we need not dwell on this potential shift of target, for my criticism of State Space Semantics (see sections 5.5, and 5.6 below) is rooted in different grounds.

contextualized *sequential* information. Simplifying statistical techniques, such as Principal Components Analysis—henceforth abbreviated PCA; see 4.6—allowed us to identify those hidden dimensions along which important variations take place. PCA was useful because it helped us make conceptual sense of the processing by 'localizing' information in hidden state space. Thanks to PCA we could observe the temporal trajectories of the hidden patterns through state space by paying attention to those input-output correlations that make the highest contribution to the net's overall ouput behaviour. To remind the reader (see chapter 4, section 4.6, above) PCA analyzes sets of hidden activation patterns, and represents their internal correlations by showing grammatically similar sentences as following *closely resembling trajectories* in the simplified space obtained by plotting particular principal components along certain axes. So, for example, a PCA performed on a simple recurrent network's representations of the sentences 'boy who boys chase chases boy', and 'boys who boys chase chase boy' yielded the following hidden space trajectories by plotting the second principal component along the ordinate (see fig. 5.2).

Churchland (1998) considers how Laakso and Cottrell's experiments might apply to the case of simple recurrent networks. As figure 5.2 illustrates, representational similarity *within* a network consists in the spatial proximity of the trajectories obtained as an effect of the sequential processing undergone. We now only need a notion of similarity of trajectory within a hidden space between distinct recurrent networks. Extrapolating from the case of simple feedforward networks,




[Fig. 5.2]: Trajectories through state space for '[boy who boys chase chases boy' and 'boys who boys chase chase boy']. After the indicated word has been input, each point marks the position along the second principal component of hidden unit space. Magnitude of the second principal component is measured along the ordinate; time (i.e., order of words in sentence) is

measured along the abscissa. [...The] sentence-final word is marked with a ]s. (Adapted from Elman, 1992, pp. 162-3)

Churchland contends:

> Two networks have the same conceptual organization if and only if there is some rotation, translation, and/or mirror inversion of the prototype-trajectory family of the first network such that, when the space (or relevant subspace) of the first network is projected onto the space (or relevant subspace) of the second, all of the corresponding trajectories (as identified by what sensory inputs activate them) coincide perfectly. (Churchland, 1998, p. 29)[8]

With this criterion at hand, Churchland reaffirms the credentials of State Space Semantics:

> The account we are currently piecing together ... is not just a syntactic account; for it promises to do what we have always expected a semantic theory to do. It ... provides a criterion for assigning the *same* contents to the representational vehicles of distinct individuals. It gives us, that is, a criterion for accurate translation across the representational/cognitive systems of distinct individuals. (*Ibid.*, p. 31)

Churchland (personal communication) believes that the current debate over the fate of State Space Semantics has a direct bearing upon the connectionist

---

[8] Strictly speaking we may need to compare *actual* trajectories, rather than *prototypical* ones, for the latter are abstractions generated statistically, and thus are causally inert as far as the dynamics of the processing goes (see chapter 4, section 4.7 above, and chapter 7, section 7.4 below). We may stay with prototypical trajectories, for present purposes, with the proviso that the argument can be put in terms of actual trajectories, at the expense of having to compute 'many' more distance relations.

defence of Quine's Inscrutability Thesis produced in chapter 4. In particular, Churchland claims that a connectionist sympathizer of Wright may be able to exploit the notion of 'similarity of prototypical trajectory' spelt out on this section, in order to discredit perverse theories of semantics. As I argued in chapter 4, a Quinean may seek to naturalize concepts via a LOT hypothesis, or in a connectionist architecture. In the former case, our standard theory, ST, would be favoured over the fully-perverse alternative, $PT_4$ (see section 4.2 above). Whereas in the latter case, a connectionist setting, I argued, is neutral between ST, and $PT_4$. Churchland, however, believes that the foe of Quine is not forced to endorse, as my twofold picture suggests, a LOT hypothesis in order to make her case. A connectionist sympathizer of Wright may well go for the second option, while denying the alleged neutrality between the standard and the perverse renderings of Native. Churchland speculates that whereas we should expect to find that the prototypical trajectories of Native sentences coincide perfectly with the prototypical trajectories of standard English sentences, the prototypical trajectories of perverse English sentences, we may expect, will diverge, showing thus that Native and perverse English lack a common conceptual organization (the reason for this disanalogy will become apparent in the next section). If Churchland's considerations are on the right track, the foe of Quine may be able to recast Wright's notion of 'psychological simplicity' in terms of prototypical trajectories, and argue that perverse prototypical trajectories are doomed to be more complex, due to the complexity that afflicts the conceptual repertoire of the basic clauses of the perverse semantic theory (see 4.2 above).

In the next section I shall argue that the Quinean has nothing to fear from these considerations. In particular, I shall try to show that Churchland's defence of State Space Semantics fails to bring robustness to semantic discourse, and lacks a connectionist notion of synonymy of the kind required by a robust theory of translation, and by extension, by a robust theory of mental representation, which would discredit perverse semantic theorizing.

## 5.5 *A Connectionist Approach to Radical Translation: First Reply to Churchland*

For argument's sake, I will agree with Churchland's first contention—namely that fit of prototypical trajectories via rotations, translations, etc. provides us with a connectionist notion of conceptual similarity.[9] We may also agree, in virtue of Laakso and Cottrell's experimental results, that neural networks do create hidden representations whose contents can be *objectively* compared—although see sections

[9] Although it is not clear to me whether fit of prototypical trajectories via rotations, translations, etc. is a *necessary* and sufficient condition for conceptual similarity, rather than a sufficient condition–fullstop. Churchland (1998, p. 29) expresses similar worries. However, I don't think our worries are motivated by the same problem. According to Churchland, the fitting of the trajectories may be only a sufficient condition in view of cases where concept identity across individuals involves causal connections to very different environmental features. Churchland's favourite example is Isaac Newton and Christian Huygens' conceptions of light as stream of particles, and as wave train, respectively. I would simply argue that whether fit of trajectories is a necessary condition or not for conceptual similarity is purely an empirical question, independent of whether concepts across

5.6, and 5.7, below. This certainly marks a watershed with respect to a mere connectionist syntactic theory. But the question I now want to pursue is: Can State Space Semantics provide a criterion for specifically one-to-*one* translational mappings across networks? In what follows I shall introduce a connectionist reading of Quine's Thesis of the Inscrutability of Reference along the lines proposed in chapter 4 (section 4.6), in order to argue that State Space Semantics cannot provide such a robust criterion.[10]

Let us consider extensions of the three semantic theories of Native reviewed in chapters 1, and 3 (the standard theory, ST; Hookway's hybrid alternative, $PT_3$; and the fully-perverse proposal that I advanced in chapter 3, $PT_4$). The extensions of ST, $PT_3$, and $PT_4$—ST*, $PT_3$*, and $PT_4$*—are meant to produce behaviourally supported satisfaction conditions for the Native compound expressions 'blanco gavagai' and 'blanco gato'. Natives utter 'blanco gavagai' and 'blanco gato' only

---

individuals are linked to the world in similar ways or not. Fleshing out this thought would take us far afield from our present purposes—see Calvo Garzón (in preparation b).

[10] Let me stress from the start that a rebuttal of Churchland's criterion is not necessarily dependent upon agreement on Quine's Inscrutability Thesis. Parallel arguments to the one I'm about to offer may well be urged by an anti-Quinean (thanks to an anonymous referee for *Philosophical Psychology* for stressing this point in response to a previous version of this chapter—see Calvo Garzón, 2000b). Nonetheless, although focusing on the theory of *reference* from a Quinean perspective simply shows my personal biases, that will permit me tackle straightforwardly a potential line of attack to be reviewed in due course against the perverse semantic theory of reference I offered in chapter 3, and defended in chapter 4. For an overall attack on Churchland's general theory of *content* not dependent on semantic skepticism as prompted by Quine see Calvo Garzón (in preparation b).

when a white rabbit, and a white cat show up in their visual field, respectively. Our standard theory, ST*, would deal with the satisfaction conditions of those expressions in the following way:

**ST***

Axioms:

**(a)** (x)(x satisfies 'gavagai' iff x is a rabbit)

**($a_1$)** (x)(x satisfies 'gato' iff x is a cat)

**($a_2$)** (x)(x satisfies 'blanco'^$f$ iff (x is white & x satisfies $f$))

Theorems:

**($a_3$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is a rabbit))

**($a_4$)** (x)(x satisfies 'blanco'^'gato' iff (x is white & x is a cat))

On the other hand, an extension of Hookway's disjunctive route (see chapter 1, section 1.6, above) would account for the Native compounds as follows:

**$PT_3$***

Axioms:

**(b)** (x)(x satisfies 'gavagai' iff x is an undetached rabbit part)

**($b_1$)** (x)(x satisfies 'gato' iff x is a cat)

**($b_2$)** (x)(x satisfies 'blanco' iff either

(a) 'blanco' occurs together with 'gavagai' and x is an undetached part of a white rabbit

or

(b) 'blanco' occurs in some other context and x is white)

Theorems:

**($b_3$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is an undetached part of a white rabbit))

**($b_4$)** (x)(x satisfies 'gato'^'blanco' iff (x is white & x is a cat))

And finally, an extension of the fully-perverse semantic theory $PT_4$, $PT_4$*, would offer the following counterpart:

**$PT_3$***

Axioms:

**(c)** (x)(x satisfies 'gavagai' iff x is a 99% undetached rabbit part)

**($c_1$)** (x)(x satisfies 'gato' iff x is a 99% undetached cat part)

**($c_2$)** (x)(x satisfies 'blanco'^*f* iff (x is white & x satisfies *f*))

Theorems:

**($c_3$)** (x)(x satisfies 'blanco'^'gavagai' iff (x is white & x is a 99% undetached rabbit part))

**($c_4$)** (x)(x satisfies 'blanco'^'gato' iff (x is white & x is a 99% undetached cat part))

Assuming that ST* is behaviourally fully adequate, $PT_3$* and $PT_4$* are behaviourally fully adequate too. A translator guided by either $PT_3$* or $PT_4$* will predict native assent to/dissent from the queries 'Blanco gavagai?' and 'Blanco gato?' in exactly the same sort of circumstances in which one guided by ST* would.

Imagine now that we train a simple recurrent network, call it N, on Native sentences of which these are examples:

**(1)'** Blanco gavagai.

**(2)'** Blanco gato.

Also we train three simple recurrent networks—call them network A, network B and network C—with English sentences derived from ST*, $PT_3$* and $PT_4$*, respectively.[11] Sentences for network A are:

**(1)** There is a white rabbit.

**(2)** There is a white cat.

Network B's counterparts are:

**(1)*** There is an undetached part of a white rabbit.

**(2)*** There is a white cat.

And, finally, sentences for network C are:

**(1)**** There is a white 99% undetached rabbit part.

**(2)**** There is a white 99% undetached cat part.

According to Churchland's earlier conclusion (see section 5.4), State Space Semantics should furnish us with an objective criterion for judging sameness of content which will deliver an accurate translational map.[12] Imagine then a thought-

---

[11] This is just a thought experiment. I shall ignore the technical adjustments required in the architecture and training regime with respect to Elman's above simulation.

[12] An anonymous referee for *Philosophical Psychology* points out in response to a previos version of this chapter—see Calvo Garzón (2000b)—that the truth-conditional semantics invoked to spell out ST*, $PT_3$*, and $PT_4$* might be at odds with Churchland's connectionist approach to semantics. This, however, should not cause any concern. As noted in chapters 2, and 4, we may naturalize concepts,

experiment in which we apply Principal Components Analysis to the sentences produced by networks N, A, B, and C. Consider first just N and A. We should expect to find that the prototypical trajectories of (1) and (2) in A would bear a strong correlation in certain hyperplanes, as identified by Principal Components Analysis (see section 5.4 above), to the prototypical trajectories of (1)' and (2)' in N respectively. Why is that the case? In Churchland's view, the driving force in assigning content to the prototypical trajectories of sentences (or for that matter, to prototypical points in feedforward networks) comes in terms of the *relative* spatial position which trajectories (or points) bear to one another *within* a representational space. In other words, content is *primarily* assigned—although see below—as a function of the concept-to-concept relations holding within a cognitive system. We may then conclude that the prototypical trajectories in N for 'blanco gavagai' and

---

going from natural languages to mental representations, by focusing upon the relation between the concepts belonging to a speaker's conceptual repertoire, expressed by words, and the information content of real internal states in her brain. So, assuming there is such a relation—and Churchland (personal communication) agrees—ST*, $PT_3$*, and $PT_4$* will each find a counterpart in State Space Semantics such that a network's representation of, say, the phrasal concept BLANCO GAVAGAI consists of a particular pattern of activation across its hidden units. In this picture, semantic content consists of a particular combination of values along each of the relevant dimensions that define the subspace in question. Thus, by following the standard semantic theory, ST*, a hidden pattern of activation $\langle h_1, ..., h_n \rangle$ across the hidden units $\{H_1, ..., H_n\}$ will carry information about *white rabbits*, as a function of the degree of rabbitness and whiteness along RABBIT and WHITE semantic dimensions. Similarly, a State Space Semantic reading of $PT_3$*, and $PT_4$* will deliver representations identifiable, along other dimensions—along UNDETACHED/RABBIT/PART, and 99%/UNDETACHED/RABBIT/PART semantic dimensions, respectively.

'blanco gato' perfectly correlate with the prototypical trajectories of sentences (1) and (2) in A. And the reason for this is that the internal relations of Native sentences are isomorphic to the internal relations that hold for "standard English" sentences: For instance, 'blanco' bears the same relation to 'gavagai' and 'gato' as 'white' does with respect to 'rabbit' and 'cat'. Following Churchland's earlier suggestion, there will be some rotation, translation and/or mirror inversion of the network A's prototypical trajectories such that they will match perfectly all trajectories obtainable in N's space.

Assuming this to be the case, next question is: Does this connectionist account of content similarity give us a one-to-one mapping between Native and English? In other words, will the isomorphism found between N and A reemerge when comparing N with B, and with C—fully perverse English? Churchland certainly does not want this to be the case, for he is willing to conclude that State Space Semantics provide us with the means of achieving a robust translation between languages, as we should expect from a rigorous theory of semantics (see Churchland, 1998, p. 31). However, I shall argue that whereas in the case of B (the hybrid theory, $PT_3$*), Churchland may be right, in the case of C (our fully perverse theory, $PT_4$*), we will find a perfect isomorphism with respect to N, or at least, as perfect as the isomorphism between N and A is supposed to be.

Under $PT_3$*, the satisfaction conditions of 'blanco' are linked to *undetached parts* of white-... when 'blanco' is coupled with 'gavagai'. In all other cases, $PT_3$* behaves standardly, taking 'blanco'-related utterances to be associated with whole

enduring white cats, for example. Hence we may predict that the relation that 'white' bears to 'rabbit' and to 'cat' in network B is a heterogeneous relation.[13] On the other hand, the relation that 'white' bears to 'rabbit' and to'cat' under network A is an homogeneous relation. And since we are assuming that the relation that 'blanco' bears to 'gavagai' and 'gato' is homogeneous as well, the prototypical trajectories in network B's hidden space will diverge, we may predict, with respect to the trajectories obtained in the Native Network.[14] Nevertheless, the hybrid character of $PT_3$* (i.e., standard-cum-perverse) seems to be alien to Quine's original pursuit. Quine's aim was to produce a *fully perverse* alternative to ST in the sense that for *every* standard referent that ST picks out, a perverse counterpart is offered. This is precisely what $PT_4$* achieves.

---

[13] The reader not familiar is urged to visit chapters 1, and 2 for an appraisal of Hookway's 'divide-and-rule' perverse semantic strategy.

[14] The reader may wonder whether we could broaden the scope of $PT_3$*'s perversity. Axiom ($b_2$) in $PT_3$* would then need to have indefinitely many disjuncts (see chapter 2, section 2.1 above). We will require an indefinite number of disjuncts in order to link the satisfaction conditions of 'blanco' to the appropriate wholes of undetached parts of rabbits, cats, cows, paper, etc., etc. And the same would happen with respect to all those axioms required for dealing with any other Native colour-word. Therefore, it *may* be the case that the perverse semanticist will not be able to state a fully-perverse *disjunctive* semantic theory. However, we ought to notice that this difficulty is rooted on rather speculative grounds. As I noted in chapter 2, it is not obvious that the aforementioned difficulty could not be overcome by some baroque plot which the Quinean has up his sleeve. Nevertheless, I shall not expand on these considerations, for if we were able to *homogenize* the internal relations of $PT_3$*, we would have a perfect isomorphism with respect to Native, which is what I aim to show now with $PT_4$*.

Under our fully perverse network C, the relation that 'white' bears to 'rabbit' and 'cat' is an homogeneous relation. The relation that 'white' bears to 'rabbit' and 'cat' under network C is exactly the same *internal* relation as the one that 'white' bears to 'rabbit' and 'cat' in network A.[15] We supposed above that the internal relation 'white' bears to 'rabbit' and 'cat' in A is the same internal relation as 'blanco' bears to 'gavagai' and 'gato' in N. Therefore, prototypical trajectories in network C's hidden space will be similar to the prototypical trajectories in N. That is, by rotating or translating the prototypical trajectories of sentences (1)** and (2)**, we'll find that they coincide perfectly with the trajectories followed by (1)' and (2)' in N. This neatly shows, I believe, that there are no grounds for favouring sentences (1) and (2) over sentences (1)** and (2)** as giving the semantic contents of (1)' and (2)' [16]. In this way, I conclude, a connectionist foe of Quine won't be able to discredit perverse semantic theorizing as suggested by Churchland (see 5.4 above). Granting that prototypical trajectories of standard English sentences coincide perfectly with those of Native sentences, there are no grounds—or, at least, no grounds revealed by the light of Churchland's considerations—for maintaining that perverse English trajectories won't fit equally well.

---

[15] Note that derivations in PT have exactly the same syntactic structure as derivations in the standard theory, ST (see chapter 4, sections 4.6, and 4.8 above).

[16] In response to previous versions of this chapter, some philosophers have worried that considerations regarding simplicity, both in the axiomatic and derivational structure of semantic theories, and in the psychological theory that accompanies semantic theorizing, could discredit $PT_4$*. For arguments against structural and psychological simplicity constraints, see chapters 2, and 4, respectively.

In the remainder of this section I shall address a potential rejoinder that someone sympathetic to Churchland may try out. But before that let me introduce a caveat to deal with a potential source of misunderstanding. Someone may worry that the argument I've advanced in this section relies too heavily on the internalist part of Churchland's theory of content.[17] As I mentioned above, Churchland's way of determining content comes *primarily* in terms of the internal similarity among prototype-trajectories. In simple cases as the toy languages we've been considering, Churchland would agree that we can safely put the burden on the internalist side—Churchland (1998, pp. 29-30). However, not all constraints on content assignment are going to be internal—and so Churchland agrees (see section 5.2 above). We need to consider the *external* causal relations linking trajectories and points in hidden space to environmental features. Someone might then hope that we may be able to exploit some sort of *externalist constraint* to 'anchor' content, bringing, thus, robustness to semantic theory. I believe that this putative line of argument is doomed. Fortunately, having developed my argument by looking at Quine's parable of Radical Translation—see fn. 10 above—it won't be difficult to see why.

The externalist part of Churchland's theory of content would highlight the fact that networks A, B, and C stand in different causal relations to "stable and objective macrofeatures of the external environment" (see Churchland, 1998, p. 8). Nevertheless, even though different networks may enjoy orthogonal patterns of connectivity with the environment, the very point of Quine's Inscrutability Thesis is

---

[17] Thanks to an anonymous referee of *Philosophical Psychology* for bringing this worry to my attention—see Calvo Garzón (2000b).

that there is *no fact of the matter* as to which objective macrofeatures are the ones being pinned down—see chapter 1. Churchland seems to ignore this obvious point when he notes that:

> [what] we have, then, is [...] networks with highly idiosyncratic synaptic connections; [...] networks with hidden-layer neurons of quite different microcontents; [...] networks whose input-output behaviors are nevertheless identical, *because* they are rooted in a common conceptual framework embodied in the activation spaces of their respective hidden layers. (Churchland, 1998, p. 11; emphasis added)

I ignore what moves Churchland to make such a strong contention.[18] We may fix the representational content of a given hidden pattern of activation by considering, partly, the causal patterns of connectivity between the input—sensory—units of the network, and those environmental macrofeatures that are responsible for the spread of activation to the hidden layers. However, since the relevant environmental features are *observationally indistinguishable* (see chapter 1 above), we cannot appeal to externalist constraints in order to single out *one* particular correct translational mapping of N—rabbits, say—as opposed to the others. This clearly illustrates a weakness in Churchland's defence. Note that the fact that different network's input-output patterns of behaviour can be identical need not come, *contra*

---

[18] Indeed, "[...] input-output behaviors are nevertheless identical, *BECAUSE* they are rooted in a common conceptual framework" (capitalization and emphasis added) seriously risks begging the whole issue in Churchland's defence of State Space Semantics. Nevertheless, for present purposes, we need not press on this point—see Calvo Garzón (in preparation b).

Churchland, as a consequence of sharing a common conceptual framework. But all this is by now, I hope, pretty obvious. Let us move on then to a more interesting line of response hinted by Churchland.

Churchland (personal communication) agrees with the general line of argument of this section. In particular he agrees that there will be some systematic isomorphism between the trajectory-structures of networks A and C—i.e., the standard and the fully-perverse networks—such that we would be justified in *pairing* the standard and the fully-perverse translations as the inscrutable alternatives. However, Churchland is not ready to surrender. And the reason is, Churchland believes, that networks A and C will display some *fine-grained* structure that hopefully can be distinguished under Principal Components Analysis.[19] Someone sympathetic to Churchland may then hope to exploit these potential fine discriminations in the following way: Suppose network C is trained to achieve grammatical competence on an extended set of fully-perverse sentences, which will require it to master, among other things, the grammar of *percentile fractions*, the grammar of *wholes* and *parts*, both *detached* and *undetached*, and a substantial *vocabulary* that is *absent in the coding activity of network A*. We may therefore be able to discriminate between the two networks by examining their

---

[19] In fairness to Churchland it must be noted that the worry I am about to introduce next is not fully worked out, but is a preliminary reaction of Churchland to a previous version of this paper. Since the line of argument is not fully developed, it will be difficult to submit to critical scrutiny. We may then read the remainder of this section as a sketched worry prompted by a hypothetical sympathizer of Churchland.

respective state-space trajectories. Prototypical trajectories in network C will presumably have additional 'kinks' and 'elbows', which will reflect the additional words whose combinations make up those trajectories. This, despite the fact that its *coarse*-grained structure might map up rather nicely onto the prototypical trajectories of network A.

I believe, however, that in our present case, this putative line of response is also doomed to failure. The reason is that we are to assume that Elman's model can be extended to encompass the processing of a real natural language. If it can then there is *no* vocabulary deployed by network C that is absent in the coding activity of network A. Fortunately, the discussion in chapter 4 will help us see the reason for this. As I argued in section 4.2, we may devise translation manuals for fellow speakers of our Home language (see chapter 4, fn. 4). Therefore, a propos the standard and the fully-perverse training domains of section 5.5 (networks A, and C), I may translate your English sentence 'There is a white rabbit' homophonically as my 'There is a white rabbit'. Or I could translate it heterophonically as my 'There is a white 99%-urp'. Since my sentence 'There is a white 99%-urp' is a well-formed sentence of English, it is one you could produce and, hence, must be subject to translation into my English. Again, my homophonic manual would equate it with my 'There is a white 99%-urp', whereas my heterophonic manual would translate it as 'There is a white 99% undetached part of a 99%-urp'. Once again, this sentence is also a well-formed sentence in your English. So, once again, I can translate it either via my standard manual or via my perverse manual. Obviously the process iterates indefinitely. This neatly illustrates the fact that whatever vocabulary the

enlarged network C deploys will also be present in the coding activity of the enlarged network A. In short, in the enlarged case the languages that networks A and C are trained on are formed out of the same lexicon. I conjecture that we won't then be able to discriminate between them by looking at additional 'kinks' and 'elbows' in their respective trajectories since, even though we may build increasingly complex phrasal structures by the usual combinatorial means, these structures belong to the same lexical body, *and* enjoy similar internal relations within each network. If these considerations are correct, then it follows that Churchland cannot appeal to fine-grained divergencies to make his case. I conclude, *contra* Churchland, that State Space Semantics does not provide a robust criterion for accurate translation across individuals; and having developed my argument by looking at Quine's Inscrutability Thesis as illustrating indeterminacy in the semantic field, the conclusion to draw is that State Space Semantics is not a viable candidate to exemplify robustness across representational/cognitive individuals—*pace* Churchland, 1998, p. 31.[20]

## 5.6 *The Collateral Information Challenge: Second Reply to Churchland*

[20] It must be stressed, in fairness to Churchland, that the issue won't be settled purely on theoretical grounds. In Calvo Garzón (in preparation a) my goal is to see if Churchland's claim can be falsified empirically by training networks A, B, and C on different sets of sentences derived from ST*, $PT_3$*, and $PT_4$*, respectively, and computing the correlations of trajectories across networks. I hope that these neurosimulations will back up the theoretical argument of this section.

In sections 5.2-5.5, I focused exclusively upon one part of Fodor and Lepore's attack against State Space Semantics. Namely, the part where Fodor and Lepore exploit the architectural and functional idiosyncracies of connectionist networks in order to weaken the metaphysical status of State Space Semantics. We saw how Churchland's deployment of Laakso and Cottrell's neurocomputational results failed to bring the required robustness to connectionist semantics. In this section I'd like to draw the reader's attention to a different problem for Laakso and Cottrell's rebuttal of Fodor and Lepore's criticism, and Churchland's subsequent defence of State Space Semantics; a problem that to the best of my knowledge has been completely ignored in the connectionist literature.[21] Fodor and Lepore exploit the potentially orthogonal histories of different individuals to introduce what they labeled 'the collateral information problem':

> The point is that if a semantics recognizes dimensions of state space corresponding to all the properties of dogs about which our beliefs differ, then even assuming that your state space has exactly the same dimensions as mine, the location of the dog concepts in our respective spaces is likely to turn out to be quite significantly different. This should be all sounding like old news;

---

[21] Although this other problem is not directly relevant to the connectionist defence of the Inscrutability Thesis advanced in chapters 4, and 5, I considered that it would be of interest to the general reader to discuss it before ending the first part of my dissertation. This is in part due to the lack of echo in the literature, and the fact that Laakso (personal communication), Cottrell (personal communication), and Churchland (personal communication) have acknowledged that Fodor and Lepore's forthcoming criticism has not been addressed in their respectives replies—Laakso and Cottrell (1998; 2000), and Churchland (1998).

> it's just the worry, familiar from attempts to construct a notion of content identity, that a lot of what anybody knows about dogs counts as idiosyncratic; it's "collateral information", the sort of thing that Frege says belongs to psychology rather than semantics. If we are to have a notion of meanings as shared and public property, a robust notion of meaning, we must somehow abstract from this idiosyncratic variation. (Fodor and Lepore, 1996, pp. 156-7)

In this passage, Fodor and Lepore target Churchland's earlier reading of State Space Semantics (that is, Churchland's definition of semantic similarity in terms of *absolute* positions in state space with respect to a given set of dimensions; see section 5.2)[22] We may update their charge to address Churchland's latest approach to connectionist semantics (that is, Churchland's definition of semantic similarity in terms of *relative* positions in state space, such that different dimensionality across networks becomes harmless—see section 5.2). Their charge becomes: Even assuming that the set of relations among the patterns of activation in your state space is similar to the set of relations obtained in mine, and even assuming that the distances among the sets of vectorial representations in our respective state spaces are highly correlated, the location of the dog concepts in our respective spaces may still differ significantly.

Fodor and Lepore exploit the fact that different individuals are likely to have had very different encounters with diverse environmental features. The reader, nevertheless, may wonder why this should be a problem. As Fodor and Lepore (1996, pp. 157-ff.) point out, it is at first sight difficult to appraise how we can

---

[22] Note that Fodor and Lepore assume, as a best-case scenario for Churchland, that different state spaces enjoy the same dimensionality.

entertain the same concepts if, according to State Space Semantics, *all* dimensions in hidden space determine the semantic content of our conceptual repertoires. Fodor and Lepore consider two ways out for Churchland: On the one hand, the defender of State Space Semantics may help herself to an analytic/syntehtic distinction, in order to discriminate between those hidden dimensions that are highly relevant in determining content—and which hopefully we all share!—and those dimensions which are less relevant, or not relevant at all—and which hopefully correspond to those axes in state space that reflect historical idiosyncracies. On the other hand, Fodor and Lepore argue, we may appeal to the empiricist assumption that all concepts are (statistical) functions of 'sensory' concepts. This would also furnish us with a robust account of conceptual similarity since *all* dimensions would then correspond to sensory properties. Hence, we may say that two individuals share their conceptual repertoires if they have relevantly similar sensory connections. Unfortunately, Fodor and Lepore would recommend neither of these two options to their connectionist enemy. Regarding the first option, honouring an analytic/synthetic distinction may bring well-known problems that the reader familiar with the philosophy of language literature will be aware of.[23] On the other hand, Fodor and Lepore wouldn't recommend the second option either. Although dressed in connectionist clothing—viz., statistical, rather than boolean functions—

---

[23] Treating these problems would take us far afield. However, since Churchland himself rejects the analytic/synthetic distinction, we may for present purposes agree with Fodor and Lepore, and Churchland, and ignore that option. The reader not familiar witrh the problems with analyticity may care to consult Quine's seminal 'Two Dogmas', and Haack, 1979.

the assumption is an embarrasing one that would inherit all the problems that afflicted classical empiricism. In short, State Space Semantics is caught on either of two horns—honouring an analytic/synthetic distinction, or resurrecting empiricism; and neither of these alternatives is very attractive, both in Fodor and Lepore's, and in Churchland's view.

In his latest defence of State Space Semantics, Churchland (1998) does not address the collateral information problem.[24] Laakso and Cottrell's neurosimulations showed us that conceptual similarity could be objectively measured "regardless of how inputs are encoded, and regardless of number of hidden units" (Laakso and Cottrell, 1998, pp. 595-6; see section 5.3 above). This however, I claim, does not address the collateral information problem since in Laakso and Cottrell's simulations information across networks was never compared *regardless of the training histories of the networks*.[25] In the experiment reported in section 5.3, several networks were trained on a colour-categorization task. The networks were trained on four different encodings of the input data: a *real* encoding, a *binary* encoding, a *gaussian* encoding, and a *sequential* encoding—see

---

[24] It may be the case that Churchland drops the issue after his exchanges with Fodor and Lepore in McCauley (1996), confident that the battle has been won. I doubt that this is the case, but I won't press on this issue here (see Calvo Garzón, in preparation b). The purpose of this section is more modest. I shall simply argue that Churchland wouldn't be able to appeal to Laakso and Cottrell's results to address the 'collateral information problem'. I encourage the reader to consult McCauley (1996) to check by herself whether Churchland is justified in dropping the issue.

[25] Cottrell (personal communication) emphasizes that their neurosimulations were never intended to address this problem, and that they actually shouldn't—see below.

Laakso and Cottrell, 1998, for the details. All encodings, however, were variations of the *same* set of data. The idea was to illustrate that individuals with contrasting sensory modalities may categorize the world in similar ways.

If we are to make use of Laakso and Cottrell's neurosimulations to address the collateral information problem, I contend, we need to train different networks on *different sets of data* (that would amount to equipping individuals with different histories of categorization and concept acquisition). For purposes of illustration, let us consider the following two ways in which this may be accomplished. On the one hand, networks may be trained under what I shall call an *inductively robust* training regime. An inductively robust training regime comprehends a set of data that allows a network to induce a given regularity with the employment of the (ideally) minimum number of samples (see below). On the other hand, by *inductively weak* training regime I have in mind a set of data such that a network being trained on this set requires to 'see' a large number of samples before being able to induce the same regularity from the environment.[26]

It will be easier to illustrate the distinction with a simple example: Imagine a simple feedforward network trained to perform an addition function, such that the

---

[26] It goes without saying that the input patterns presented to a pair of networks under inductively weak, and inductively strong training conditions, respectively, must belong to the same base data set. Notice that no correct generalization can be learned outside a given training space—cf. Elman, 1998. The dubbing ('inductively robust' *versus* 'inductively weak' training regimes) is due to Bill Casebeer. Many thanks to him for very helpful discussion on the topic of this section.

sum of the input values gives you the output value.[27] This can be achieved simply by designing a network with two input units, two hidden units, and one output unit (with a linear activation function—see chapter 4, section 4.4). For purposes of illustration, let us consider only inputs between 0.0 and 0.5 with a single decimal place, such that there are 36 possible sets of input patterns, and the output value ranges between 0 and 1. Although the network could be presented with 36 combinations of members, the network need not see all of them in order to master its task. In fact, a network exposed to just 5 sets of exemplars can successfully learn the addition function for all 36 possible cases. The trick is simply to expose the network to a representative portion of its overall training domain. Thus, the following sets of input patterns would constitute what I called earlier an inductively robust training regime:

**Training regime a**

<0.0, 0.0>,
<0.1, 0.2>,
<0.3, 0.3>,
<0.4, 0.4>, and
<0.5, 0.5>.

**Training regime b**

<0.0, 0.0>,
<0.1, 0.1>,
<0.2, 0.2>,

[27] The example is taken from Plunkett and Elman (1997), although their discussion of the network is alien to our current concerns.

<0.3, 0.4>, and
<0.5, 0.5>.

On the other hand, the less representative the samples are (i.e., the more restricted they are to specific regions of the training domain), the bigger the amount of data the network will have to see in order to induce the correct generalization. In this way, the follwing set of input patterns would constitute what I called earlier an inductively weak training regime:

**Training regime c**
<0.0, 0.2>, <0.0, 0.3>,
<0.1, 0.2>, <0.1, 0.3>,
<0.2, 0.0>, <0.3, 0.0>,
<0.4, 0.0>, <0.0, 0.4>,
<0.1, 0.1>, <0.2, 0.1>,
<0.3, 0.1>, <0.2, 0.2>,
<0.5, 0.5>, and <0.0, 0.0>.

Notice that, unlike in training regimes a, and b, where the subset chosen spans the full range of possible outputs at evenly-spaced intervals, training regime c draws most of it examples from cases where the sum of both inputs is between 0.2, and 0.4. A network trained under these conditions is exposed to less representative samples, and will thus require to see many more patterns before being able to

induce the same regularity discovered by networks trained under 'robust conditions' a, and b, above. [28]

I shall argue next that Laakso and Cottrell's technique (5.3) can deliver high correlations *only* when measuring similarities across networks that have been trained on inductively robust regimes. Unfortunately for the friend of State Space Semantics, to address the problem of collateral information we need to compare networks that have been trained under inductively robust conditions with networks trained under inductively weak conditions, I contend.[29] Let me elaborate.

Following the above distinction between inductively robust *versus* inductively weak training regimes, we may interpret the collateral information problem in at least two ways. In a best-case scenario for Churchland, we may identify the histories of two individuals with two networks both trained under different, although inductively robust, conditions. In that case, even though the networks are exposed to different input patterns, they will partition their state spaces similarily by sampling, for example the base data set at different evenly-spaced intervals (as training regimes a, and b, above exemplify). A high degree of correlation would

---

[28] Obviously, if the distribution of data is not representative at all of the full range of possibilities, the network will learn a somewhat different function. For argument's sake I assume that the network ends up below a pre-established area in the error landscape, so that it adds correctly any two pairs of numbers between 0.0 and 0.5. There are further technical subleties that we may obviate for present purposes that I discuss elsewhere—see Calvo Garzón, in preparation b.

[29] To be precise, comparing networks *both* being trained under diverse inductively weak regimes would furnish a more realistic setting to address Fodor and Lepore's challenge. Although to make the point more vivid, I shall compare inductively robust with inductively weak training regimes.

then be expected. Given that the different training sets span the full range of possible outputs, we may expect the set of internal relations among points in one hidden space to be isomorphic with the set of internal relations in the other network's hidden space. In a worse-case scenario, however (although see fn. 29 above), one network would be trained in an inductively robust regime, and the other network in an inductively weak regime. These networks will be trained on input patterns which are, respectively, more and less representative of their common task.[30] I take it that this is a more plausible interpretation of the collateral information problem. After all, my concept dog and your concept dog may plausibly be associated with highly divergent inferential regimes—think of me as a dog breeder (an inductively robust training environment), and you having had spare encounters with dogs in your life (an inductively weak training environment) who has eventually come to be able to tell dogs from non-dogs as well as I can. Unfortunately, under this second scenario (inductively robust *versus* inductively weak training regimes), we will expect to find a low correlation across our

---

[30] Someone may wonder whether *any* network trained under inductively weak conditions will be able to extract the correct generalizations. This should not be a problem. Hornick *et al.* (1989) have established that a simple feedforward network with a sigmoid activation function can behave as a universal approximator. That is, for any given function with a finite range and a finite domain, the function can be computed to an arbitrary level of accuracy. A different point is what counts as an arbitrary level of accuracy. Addressing this issue would require introducing some conceptual machinery in connectionist theory, such as 'error landscapes', that would take us far afield. The point, however, is too technical to be of import to this chapter—although see Calvo Garzón, in preparation c.

networks. Networks trained under inductively different conditions will be exposed to different, and maybe orthogonal, experiences on input pools of data—while acknowledging the fact that a public concept is being shared. Assuming that Fodor and Lepore's challenge is to be read in this way, different networks are highly unlikely to come out with similar solutions, or to partition their state spaces in similar ways (cosmic flukes apart!). The set of internal relations among points in one hidden space won't be isomorphic with the set of internal relations in another network's hidden space.

I conclude that a connectionist sympathiser of Churchland could not make use of Laakso and Cottrell's neurosimulations to address Fodor and Lepore's collateral information challenge. Obviously, the above considerations are not unsurmountable, and, in fairness to Churchland, the case against State Space Semantics is less than conclusive. For one thing, someone may be ready to bite one or other of Fodor and Lepore's bullets, and grant either an analytic/synthetic distinction, or empiricism. However, since Churchland explicitely rejects both options, I didn't consider them in the present discussion.

## 5.7 *Conclusion*

In this chapter, I have argued that appealing to Laakso and Cottrell's mathematical measure of conceptual similarity does not bring Churchland's optimistic conclusion. Namely, the conclusion that State Space Semantics can furnish us with a determinate theory of semantics and a robust theory of translation. Wrapping up

their overall criticism, Fodor and Lepore (1996) argue that State Space Semantics looks pretty much like an updated version of empiricism, with all its flaws. Churchland (1998), and Laakso and Cottrell (1998; 2000) argue that State Space Semantics, when reinforced with Laakso and Cottrell's results, can be distanced from empiricism: Conceptual similarities in hidden space can be objectively measured regardless of idiosyncrasies at the level of the input encoding. Churchland has ironically urged that if we are going to start with historical comparisons, his proposed connectionist theory fits better with Platonism. The moral of this chapter is neither Hume, nor Plato; Connectionist Semantics provides the right tool kit for a "connectionist Quinean" to kick in with his old-fashioned behaviouristic arguments for the Inscrutability Thesis transposed into a neuroscientific fashion—see chapter 4 above.

On the other hand, it is worth pointing out that if a semantic *irrealist* à la Quine can make connectionism her home, the results of this chapter might have a broader impact than I have argued for here. For strategical reasons, I've assumed throughout the last two chapters a representationalist framework. Both Fodor and Lepore, and Churchland would agree that a general theory of mental *representation* is required in order to explain human higher cognitive capacities. Their disagreement *reduces* to which model of cognition is correct: A LOT model with classical constituency, and classical processing, or a connectionist model where constituency and processing are non-classical—see Calvo Garzón, 2000a. This however may prove to be a trivial distinction, were the Quinean to earn her keep as a connectionist, for *both* Fodor and Lepore, and Churchland may well sink together

in the boat of *representationalism*. But I shall leave these matters for another occasion.[31]

This chapter closes the connectionist defence of Quine's Thesis of the Inscrutability of Reference that has been the subject matter of Part I of my dissertation. The connectionist setting developed in chapters 4, and 5, however, has in my view implications for the philosophy of language, and the philosophy of mind, that go well beyond the treatment of the semantic notion of reference discussed so far. In Part II of this work (chapters 6, and 7, below) I shall propose to extend the results of chapters 4, and 5 to the debate over the ontological status of the propositional attitudes. The object of Part II will be to produce a Connectionist Defence of the Elimination of the Mental.

---

[31] In Calvo Garzón, in preparation c, I argue that a connectionist model of cognition may show that representationalist theories of mind cannot earn their keep. For some anti-representationalist positions see Keijzer, 1998; Ramsey, 1997; and Van Gelder, 1995; 1998. Notorious connectionist dissenters include Clark and Toribio, 1994; and Clark, 1997, chapter 8.

# Part II

## *The Elimination of the Mental*

---

*[Funes], no lo olvidemos, era casi incapaz de ideas generales, platónicas. No sólo le costaba comprender que el símbolo genérico* perro *abarcara tantos individuos dispares de diversos tamaños y diversa forma; [...] Su propia cara en el espejo, sus propias manos, lo sorprendían cada vez. [...] Sospecho, sin embargo, que no era muy capaz de pensar. Pensar es olvidar diferencias, es generalizar, abstraer. En el abarrotado mundo de Funes no había sino detalles, casi immediatos.*

*[Funes] was, let us not forget, almost incapable of general, platonic ideas. It was not only difficult for him that the general term* dog *embraced so many unlike specimen of different size and different forms; [...] His own face in the mirror, his own hands surprised him on every occasion [...] I think that he was not very capable of thoughts. To think is to forget a difference, to generalise, to abstract. In the overly replete world of Funes, there was nothing but details, almost contiguous details.*

—Jorge Luis Borges, *Ficciones*

# 6

## *CAN WE TURN A BLIND EYE TO ELIMINATIVISM?*

### 6.1 *Introduction*

In this chapter I shall rejoin to two arguments that Stephen Stich has recently put forward against the thesis of eliminative materialism. In a nutshell, Stich (1990; 1991) argues that (i) the thesis of eliminative materialism, according to which propositional attitudes don't exist (see chapter 7, below), is neither true nor false, and that (ii) even if it were true, that would be philosophically uninteresting. To support (i) and (ii) Stich relies on two premises: (a) that the job of a theory of reference is to make explicit the tacit theory of reference which underlies our intuitions about the notion of reference itself; and (b) that our intuitive notion of reference is a highly idiosyncratic one. In this chapter I shall address Stich's anti-

eliminativist claims (i) and (ii). I shall argue that even if we agreed with premises (a) and (b), that would lend no support whatsoever for (i) and (ii).

Before getting started, let me briefly outline the programme of this chapter. In section 6.2 I shall review Stich's first anti-eliminativist argument. Stich interprets the thesis of eliminativism as the claim that the theoretical terms of folk psychology fail to refer. Assuming that the job of a theory of reference is to make explicit the tacit theory of reference which underlies our intuitions about the notion of reference itself, Stich's argument rests upon an empirical claim. Namely, that people who know folk psychology has been discredited actually lack clear intuitions about the reference of the theoretical terms of folk psychology. Stich concludes then that the thesis of eliminativism lacks determinate truth conditions, and is not true but not false either. In section 6.3 I shall argue that Stich's anti-eliminativist conclusion does not follow from the empirical assumption he relies on. Even though intutions can be partly relevant when assigning truth values to sentences, I shall argue that an analysis of the logical form of belief-sentences still drives us to the eliminativist's conclusion. In section 6.4 I shall review Stich's second anti-eliminativist argument. What's at stake, Stich claims in opposition to his previous argument, is not whether the thesis of eliminativism lacks determinate truth conditions. Even if the eliminativist thesis were true, Stich now contends, that would be philosophically uninteresting. Stich sees the notion of reference as ultimately an *idiosyncratic* word-to-world semantic mapping. The idiosyncracy of reference, Stich believes, is what makes the eliminativist thesis philosophically uninteresting. Even though under a theory of reference, eliminativism may obtain, by assuming the idiosyncracy of

reference we are likely to encounter alternative theories of reference under which eliminativism does not follow. In section 6.5 I shall argue that Stich is exploiting the idiosyncracy of reference to give a free ride to his anti-eliminativist conclusion. To advance the flavour of my rejoinder, I shall contend that if according to our sanctioned theory of reference eliminativism follows—premise that Stich grants—, then it will still follow for any alternative theory of reference Stich may *properly* consider (although see below). Conclusions will follow in section 6.6. In addition, I shall outline Stich's (1996) latest anti-eliminativist view according to which the theory of reference just isn't the place to go to when trying to settle ontological disputes. Stich's latest twist in the eliminativist plot distances him from the arguments to be addressed in chapter 6. However, full treatment of Stich's lastest anti-eliminativist arguments will have to await until chapter 7, where I shall produce a connectionist defence of the thesis of eliminative materialism. So, without further ado, let us take Stich's arguments in turn.

## 6.2 *Eliminativism and Folk Semantics*

Stich interprets the thesis of eliminativism as the claim that the theoretical terms of folk psychology fail to refer.[1] However, Stich holds that the theory of reference is a branch of psychology:

---

[1] This is in my opinion—and according to Stich's latest views on the matter (see 6.6 below, and chapter 7, section 7.5)—a highly controversial assumption. Put bluntly, it strikes me as bizarre that ontological matters—i.e., whether something exists in the (physical world)—depends on what

> It is my contention that a 'philosophical' theory of reference is in fact a bit of psychology. It's aim is to make explicit the tacit theory of reference that is presumed to underlie our intuitions about questions like [(a) Does '____' refer to ____? and (b) Does '____' refer to anything at all?]. (Adapted from Stich, 1991, pp. 240-1)

Stich then contends that the relevant responses of psychological subjects fail to determine whether or not the theoretical terms of folk psychology refer, once they realise that folk psychology has been discredited empirically.

> [There] is good reason to suppose our tacit theory of reference says little or nothing about questions like [(a) and (b)] when the term in question is a theoretical term in a largely discredited theory. (*Ibid.*, p. 241)

Stich supports this claim by considering our commonsense *intuitions* about reference in the following experiment:

> Start with a theory that you take to be largely correct, and focus on some theoretical term central to that theory. Now imagine that the theory is found to be much worse than you supposed. One tenet after another must be rejected and replaced by a very different, and incompatible, tenet. At each step, ask whether the term, as it was embedded in the old theory, can plausibly be said

---

semantic relations a theory of reference posits. In short, it seems to me clearly intuitive that whether a relation of reference, between a set of theoretical terms and the referents they putatively pick out, obtains or fails to obtain is an *a posteriori* consequence of our ontological commitments; commitments which must be rooted *elsewhere*—see chapter 7 below. I shall, nonetheless grant Stich's interpretation without further ado for present purposes.

> to refer to anything. It is my experience that most people who play this game report that when the theory is imagined to be seriously mistaken, they often no longer have any clear intuitions about the reference of the term. (*Ibid.*, p. 240)[2]

In short, Stich's first argument against the thesis of eliminativism rests upon an empirical claim. Namely, that people who know folk psychology has been discredited actually lack clear intuitions about the reference of the theoretical terms of folk psychology. Our tacit theory of reference remains silent when confronted with questions like "Does '____ believes that p' refer to anything at all?" (*Ibid.*, p. 241). Therefore, Stich concludes, the thesis of eliminativism lacks determinate truth conditions, and is not true but not false either.

## 6.3 *Rejoinder to Stich's First Argument*

Let me start with one caveat in order to reply to Stich's above argument. Stich's argument hinges on *what* theory of reference we employ. As we saw, by favouring a psychological approach to the theory of reference, Stich reached his anti-eliminativist conclusion. An obvious starting point for a rejoinder would then be to ask whether we are justified in making use of commonsense intuitions when dealing with semantic notions such as reference. Stich (1996) wonders what makes a theory of reference correct. He considers two accounts. On the one hand, his *folk semantic*

---

[2] See also Stich (1996, pp. 46-8) for a pilot study along the lines suggested with grad students as guinea pigs.

intuitional account according to which the role of a theory of reference is to capture the theory of reference that speakers tacitly endorse. An alternative approach to reference, Stich notes, would be a scientific theory of reference whose role is to construe word-world mappings to be employed by the empirical sciences, such that reference is determined by empirical facts regardless—maybe, orthogonally—of the intuitions of the layman. Stich dubs this aternative *proto-scientific* theory of reference. Stich argues that we cannot make use of proto-scientific theories of reference unless we have an up-and-running empirical discipline where the relational notion of reference does play an active role; and we still lack such a discipline.

> So, from a practical point of view, the only way to make progress is to concentrate on the account that views a theory of reference as an attempt to describe the intuitive reference relation, the one specified by folk semantics. (Stich, 1996, p. 46)

Those sympathetic to the proto-scientific approach may simply object that intuitions and tacit theories of reference miss what's at stake. The key question is: What is the relevant evidence to construct a theory of reference? Stich is considering intuitions in response to semantic questions (a) and (b) above, which concern theoretical terms. However, the construction of a theory of reference requires the employment of many technicalities orthogonal to the intuitions of the lay man. The issue of what the notion of reference consists of, someone may contend, is not to be settled by a folk psychological theory of reference. It is rather to be settled by a *scientific theory of reference*; in fact, by the best motivated and

best regimented theory of reference at hand. It's an open question what this scientific theory of reference is, but folk psychological intuitions, if relevant at all, are not relevant in the simple and direct way that Stich's argument supposes.[3]

Nevertheless, we may go at least *part way* with Stich for argument's sake, and assume that people's commonsense intuitions may play a direct role, and further that commonsense intuitions have little to say about the reference of the theoretical terms of largely discredited theories. As I shall argue next, Stich's anti-eliminativist conclusion does not follow from this empirical assumption.

Following Stich, we may concede that intutions are partly relevant when assigning truth values to sentences. Consider the following sentence:

**(s)** Santa Claus brings joy to children.

The reference—i.e., truth value—of (s) is determined by the reference of its constitutive parts. We have then a number of options available: If the singular term 'Santa Claus' refers to the historical character, *Saint Nicholas*, we may conclude that (s) is false. Saint Nicholas does not bring joy to children. On the other hand, if 'Santa Claus' is an empty singular term, which fails to refer, (s) is neither true nor false. And finally, if 'Santa Claus' refers to a *fictional* character (see Evans, 1982, pp. 363-6), (s) would be *true-in-the-fiction*. Clearly, depending on which *intuitions* we have about the referential relation of the singular term 'Santa Claus' to the

---

[3] Those uneasy with the employment of intuitions in the first place may consult Bickle (1993, pp. 376-ff.) for a formal construal of our ontological commitments which undermines the role allegedly played by commonsense intuitions.

objects of our universe of discourse, we will assign different truth values for (s), or no truth value at all. In this example, it is reasonable to maintain, as Stich does, that speakers' intuitions do matter. Someone, moreover, may contend that Stich's 'commonsensical' approach may well generalize to other situations. Plausibly, intuitions do matter as well—at least, to some extent—in determining the reference of sentences embeding, for example, mass terms.

Consider the mass term 'caloric' which, according to XVIII abd XIX centuries' scientists, referred to a fluid substance held in bodies which produces melting, boiling, etc. Take the sentence:

**(s')** When caloric flows into a body it produces thermal expansion.

Suppose firstly intuition tells us that the extension of 'caloric' is *not* the null class—i.e., there is at least one object in our universe of discourse which falls under the mass term 'caloric'. Then, if we accept the verdict of intuition, (s') is false. The reason is that nowadays we know that thermal expansion is produced by kinetic energy, rather than by caloric fluid. Alternatively, suppose intuition tells us that the extension of 'caloric' is the null class. And, furthermore, we accept the verdict of intuition. (s') this time will come out true. Notice that (s') can be read as:

**(s*)** $(x)(y)((x \text{ is caloric} \ \& \ x \text{ flows into } y) \rightarrow x \text{ produces thermal expansion in } y)$

Now, if the predicate 'x is caloric' has the null class as its extension, then (s*) is true. The reason is that for *any* pair of objects, *a* and *b*, in our universe of discourse we may wish to consider, such that:

((*a* is caloric & *a* flows into *b*) → *a* produces thermal expansion in *b*),

the antecedent will always be false. We may then say that (s*) is true *by default* (see below).

The eliminativist, however, need not call into question these considerations. With the above proviso in mind, the eliminativist can still argue that an analysis of the logical form of belief-sentences drives us to the eliminativist's conclusion—even though intuitions may play the role Stich assignes to them. First, consider a fast route that a sympathizer of Stich may try out to obtain Stich's conclusion. Someone may argue, for example, that the sentence

**(s'')** Tom's belief that the cat was on the mat caused him to say that the cat was on the mat,

has no determinate truth-conditions. We may interpret (s''), the suggestion would run, as referring to *Tom's belief*—i.e., to *his belief that the cat is on the mat.* It would then follow that (s'') has no determinate truth conditions, since we are treating 'Tom's belief that the cat is on the mat' as an empty singular term. Therefore, Stich's conclusion obtains.

This conclusion, however, may have been too rash. We may read the logical form of belief-sentences in terms of a quantificational structure that binds a belief-variable. Take, for example, the following folk psychological law:

**(A)** If x believes that P & Q then, *ceteris paribus*, x believes that P.

Taking P and Q to be variables for lumps of Home language, we may set out the above folk psychological law in formal notation as follows:

> (x)(P)(Q) {(∃y) (x believes y & y samesays P^&^Q) → *ceteris paribus*, (∃z) (x believes z & z samesays P)}

On this reading, (A) is, by default, trivially true—if the eliminativist is right. Notice that the antecedent is false, assuming that the eliminativist is right, whatever values we choose for 'x', 'P', and 'Q'. However, admitting that the laws of folk psychology are (trivially) true does not damage the eliminativist's position. If we now examine a particular folk psychological statement, we shall see why, and see which claims of folk psychology the eliminativist rejects as false, not, as Stich requires, *neither true nor false*. Take a particular application of (A):

**($A_1$)** (∃y) (a believes y & y samesays "the cat is on the mat and someone left the window open") → *ceteris paribus*, (∃z) (a believes z & z samesays "the cat is on the mat")

To explain a's action, the folk psychologist claims:

> (∃y) (a believes y & y samesays "the cat is on the mat and someone left the window open") → *ceteris paribus*, (∃z) (a believes z & z samesays "the cat is on the mat")
> (∃y) (a believes y & y samesays "the cat is on the mat and someone left the window open")
> ---

(∃z) (a believes z & z samesays "the cat is on the mat")

The explanatory argument is valid, and its major premise is true, the eliminativist concedes. But the minor premise and conclusion, the eliminativist claims, are simply false because they are existential claims, and there is nothing we can assign as a value of y or of z which will make either true.

The reader can now see which claims of folk psychology the eliminativist rejects as false—not, as Stich requires, neither true nor false. According to the above logical transcription, 'a's believing' is not a singular term. The sentence "a believes that the cat is on the mat" does not contain a reference to a particular belief of a, but rather is a general existencial statement with respect to *believings of z by a.* Particular folk psychological statements such as the minor premise and the conclusion in the above folk explanatory argument are false. They contain variables, bound by existential quantifiers, which range over beliefs. We are thus dealing with *relational* expressions whose reference is, not an object, but a second-level function. *Contra* Stich, the question "Does '____ believes that p' *refer to* anything at all?" has the determinate answer 'Yes'. It refers to the relation (∃y) ____ believes y & y samesays p. We don't obtain *truthlessness*, as Stich requires, for 'a believes that the cat is on the mat', but rather *falsity*, as the eliminativist claims. This conclusion, moreover, is perfectly compatible with the claim that general folk psychological laws, such as (A) above, are true. The fact that (A) is (trivially) true, I contend, is to be seen as a byproduct of the *logical* apparatus in place when spelling out folk psychological laws *formally*. In short, the fact that (A) comes out true *by default* reveals that no ontological commitment is being made, and is therefore

compatible with the eliminativist claim that particular applications of (A) contain relational expressions which do *not* fail to refer, and, thus, allow us to dissolve the alleged indeterminacy urged by Stich.[4]

## 6.4 *Eliminativism and the Idiosyncracy of Reference*

Stich is quite prepared to give up the empirical premise he relied on in his first argument. Namely, that speakers lack clear intuitions about the reference of the theoretical terms of folk psychology. What's at stake now is not whether the thesis of eliminativism lacks determinate truth conditions. Even assuming that the eliminativist thesis were true, Stich now contends, that would be philosophically uninteresting. Stich sees the notion of reference as ultimately an *idiosyncratic* word-to-world semantic mapping. The idiosyncracy of reference, as we'll see next, is what makes the eliminativist thesis philosophically uninteresting.[5]

Stich favours a 'causal-historical' theory of reference—e.g., Putnam, Kripke, etc.[6] Put bluntly, after an initial reference-fixing event, reference is transmitted along a causal-historical chain. Stich then wonders how we are to discriminate between genuine and fake referential transmissions. Since the theory of reference is

---

[4] For a different attack to Stich's first argument launched by Jackson, see Stich (1996), pp. 52-4.

[5] For an origin of Stich's notion of the idiosyncracy of reference see Godfrey-Smith (1986).

[6] For argument's sake, I shall go along with Stich and grant a 'causal-historical' approach to reference. I believe, however, that the case against 'descriptive' theories of reference is not settled

a part of the theory of psychology (see section 6.2 above), Stich notes that genuine transmissions must be those "sanctioned by intuition" (Stich, 1991, p. 242). Intuitions, nonetheless, do *not* provide us with a *homogeneous* test of how word-to-world mappings are to be transmitted from the original referential baptism onwards:[7]

> [When] one looks carefully at [...the] class of transmissions that pass this test [i.e., the test of commonsense intuition], it appears that in each category the allowable events are a mixed bag having at best a loosely knit fabric of family resemblances to tie them together. The causal chain linking my use of the name 'Rebecca' with my daughter is notably different from the one linking my use of 'water' with water. And both of these are notably different from the chain linking my use of 'quark' with quarks. What ties all these causal chains together is not any substantive property that they share. Rather, what ties them together is simply the fact that common sense intuition counts them all as reference fixing chains. (*Ibid.* pp. 242-3)

And Stich goes on:

> But if it is indeed the case that common sense groups together a heterogeneous cluster of causal chains, then obviously there are going to be lots of heterogeneous variations on the common sense theme. These

---

yet; though space prohibits me from extending on this matter. For Stich's distrust of descriptive theories of reference see Stich (1990, pp. 108-ff.; 1992, pp. 254-ff.).

[7] Stich calls into question the homogeneity of the referential baptism itself, as well as the subsequent transmissions. I'll focus on the transmissions. Nothing in my present argument hangs on ignoring the ground-fixing events.

> alternatives will depart from the cluster favored by common sense, some in minor ways and some in major ways. They will link some words, or many, to objects or extensions different from those assigned by commonsense intuition. In doing so, they will characterize alternative word-world links, which we might call REFERENCE*, REFERENCE **, REFERENCE***, and so on. (*Ibid.* p. 243)

In Stich's view, there's nothing substantially different in our favoured scheme of reference—call it REFERENCE—, as opposed to REFERENCE*, REFERENCE**, etc. apart from the fact that it is the one intuition guides us towards. REFERENCE enjoys no priviledged status over its putative alternatives since the tacit rules that according to folk semantics determine our commonsense intuitions are themselves, Stich claims, a cultural product.[8]

The bearing on the eliminativist thesis is straightforward to Stich. The fact that '____ believes that p' *refers* to nothing brings no worry, since REFERENCE is a highly idiosyncratic mapping. Other relational mappings—e.g., REFERENCE*, REFERENCE**—will pick on various word-to-world semantic relations such that '____ believes that p' does *refer** to, or *refer*** to, *something*; and we have no factual reasons to favour REFERENCE over, say, REFERENCE*, but merely

---

[8] A comparison that Stich draws between folk semantics and the theory of grammar illustrates this point: "The fact that our intuitions pick out the particular word-world relation that we call *reference* rather than one of the many others [...] is largely the result of historical accidents, in much the same way that details of the grammar of our language [...] are in large measure the result of historical accidents" (Stich, 1996, p. 50).

historical reasons.[9] In the remainder of this chapter, I shall elaborate on an argument against Stich's second argument based on the idiosyncracy of reference.

## 6.5 *Rejoinder to Stich's Second Argument*

It strikes me as surprising that Stich does not provide any *specific example* of an alternative theory of reference under which the extension of '____ believes that p' is *not* empty. Stich treats beliefs as having content in virtue of causal relations linking those beliefs to referents in the world. However, Stich points out, there are a lot of causal relations out there, such that we may assign referents to sentences in a number of ways. Nevertheless, if it is actually the case that, under a particular word-world mapping, believers* and beliefs* do exist, why doesn't Stich put an example on the table, nailing thus down the eliminativist's coffin forever? As I shall argue next Stich does not do so because he *cannot* do so. The following quote from his recent *Deconstructing the Mind* reveals where Stich's argument goes astray:

---

[9] I must confess I am a little sceptical about the role played by 'historical reasons' in bringing support to the idiosyncracy of reference. The quote from Stich in fn. 8 above may bring implicitly an answer. However, Stich's position would need to be fleshed out in more detail before submitting it to critical scrutiny. Nevertheless, I shalll not press on this point here. The discussion in Part I of my dissertation clearly shows my bias towards something similar to what Stich dubs 'the idiosyncracy of reference'. Although I suspect that my motivations for endorsing it are substantially different from Stich's. Spelling out our divergencies on this matter would take us far afield.

> On the account we have been working with, eliminativism is true if and only if '_____ is a belief' refers to nothing. Let ELIMINATIVISM* be a doctrine that is true if and only if '____ is a belief' REFERS* to nothing; let ELIMINATIVISM** be a doctrine that is true if and only if '____ is a belief' REFERS** to nothing; and so on. *Clearly, some of these ELIMINATIVISM-stars are bound to be true, while others will be false.* (Stich, 1996, p. 51; emphasis added)

The key question is: Why 'bound to'? How can Stich be certain that there is going to be a semantic mapping where '____ is a belief' REFERS*...*i*...* to something? It seems to me that Stich uses the idiosyncracy of reference to give a free ride to his anti-eliminativist conclusion. To advance the flavour of my rejoinder, I shall contend that if according to our sanctioned theory of reference eliminativism follows—premise that Stich grants—, then ELIMINATIVISM*...*i*...* will follow as well, for *any* value of *i*—i.e., for any alternative theory of reference Stich may properly consider.

Granting that under REFERENCE eliminativism follows, the eliminativist's fast route to making her case is to argue that there's just *one* correct theory of reference: Namely, REFERENCE. The anti-eliminativist challenge then comes, as we saw, from the idiosyncracy of reference. Nonetheless, Stich's move unjustifiably shifts the burden of proof to the eliminativist. The eliminativist is being indirectly forced to argue that REFERENCE is the only correct theory of reference.[10] I believe

---

[10] As the careful reader will have guessed, Part I of my dissertation clearly illustrates why I wouldn't be keen on pursuing this open possibility in the logical landscape—i.e., arguing that REFERENCE is the only correct theory of reference—see fn. 9 above.

however that this is the wrong approach to the issue. An explanation must be forthcoming from the anti-eliminativist corner as to how alternative theories of reference can deliver results orthogonal to those achieved via REFERENCE (orthogonal with regard to truth value assigments). As things stand, the onus is on Stich to tell us *how* different theories of reference can deliver different results as far as ontological considerations go, *while* remaining empirically adequate.

Since Stich does not offer any particular example, we may speculate about which alternative word-world mappings would provide him with a best-case scenario. In a Quinean fashion, for instance, we may generate an indefinite number of mappings that pick out objects and extensions different from those that REFERENCE picks out; the only constraint being preservation of *stimulus meaning*.[11] Our intuitive theory of reference, REFERENCE, axiomatizes belief-predicates as follows:

**(a)** (x) (x satisfies 'belief' iff x is a belief).

However, we may easily produce a number of Quinean alternatives, REFERENCE* and REFERENCE**, which contain respectively axioms (a*) and (a**). To wit:

**(a*)** (x) (x satisfies 'belief' iff x is a temporal stage of a belief);

[11] I'll skip the details. See chapter 1 above.

**(a**)** (x) (x satisfies 'belief' iff x is an undetached belief part); and so forth.[12]

It is of course odd to talk of a temporal stage of a belief, and of undetached belief-parts. But assuming beliefs are, if they exist, the kind of things folk psychology claims them to be—functionally discrete, semantically interpretable and causally efficacious states (see chapter 7, below)—then we may take such states to have parts and temporal stages.[13]

Unfortunately, Stich cannot make use of REFERENCE* or REFERENCE**. The reason is simply that if beliefs don't exist, then temporal stages of beliefs, or undetached parts of beliefs cannot exist either! Assuming, with Stich, that under REFERENCE ELIMINATIVISM is true, it is then difficult to see how under REFERENCE* or REFERENCE**, ELIMINATIVISM* or ELIMINATIVISM** is going to be false.

What Stich requires then is a more radical way of producing alternative theories of reference—i.e., a strategy not constrained by preservation of stimulus meaning—, such that '____ is a belief' can *refer-star* to something. However, to

---

[12] Although Stich may feel uneasy with Quine's behaviouristic setting, the connectionist rendering of Quine's views on semantics spelt out in chapters 4 and 5 furnishes us with the sort of naturalized approach that Stich would find appealing. See chapter 7 below for Stich's views on connectionism.

[13] Further qualification would be required to make this counter-intuitive view tenable. We need nonetheless not worry for present purposes since, were we to discover empirically that temporal-belief-stages and undetached-belief-parts don't exist, we wouldn't even be able to generate alternative referential mappings to REFERENCE, in which case Stich would find himself unable to exploit the idiosyncracy of reference for his purposes. The problems for Stich's position, however, run deeper, and are not dependent on agreement on the above Quinean setting (see below).

illustrate why Stich's project is doomed we need not worry about how more radical examples would run.[14] Rather, let me draw your attention to an issue that has been largely ignored in the literature on Eliminativism:

It is standardly assumed among physicalists that the debate between an eliminativist and an anti-eliminativist relies on agreement on *basic* theories such as physics. Physics allegedly has the resources to explain everything. The eliminativist wishes to eliminate folk psychology. On the other hand, the anti-eliminativist wishes to reduce folk psychology to the physical level, or reconcile the two levels in some other way. Elimination or reduction is what's at stake, agreeing thus about the priviledged status of physics as an essential part of our scientific explanations.[15] The language of Physics involves notions of reference of various kinds. Note that for example reference of observational terms is different from reference of natural kind terms, or reference of highly theoretical terms. In like vein, physics involves notions of causality and notions of explanation of various kinds. However, and this is the key point, we are not to entertain alternative theories of reference which change the notions of reference, causation and explanation of our *background*

---

[14] I have in mind for example Putnam's (1981, chapter 2) model-theoretic arguments; in particular his permutation argument—see below.

[15] Obviously, the spectrum of possibilities is much broader. Anti-eliminativists may opt for any of the non-reductive materialist options available in the market nowadays. However, as far as my present considerations go, the Stich of *The Fragmentation of Reason* is a token-identity theorist, and would thus fall within the broad region I outline here—see Stich (1990, p. 103; p. 117). Nevertheless, in *Deconstructing the Mind* Stich changes his mind on the epistemic status of physics. For present purposes, we may ignore his recent shift—see chapter 7 below.

physical theory. Now, granting this, we certainly cannot assume that the referential relation which we are holding fixed for the terms of physics (e.g., quarks) will dictate to us how the reference of '_____ believes that p' is to be fixed. We may wonder then what the appropriate referential relation for the distinctive predicates of folk psychology is. Trying to provide an answer goes beyond the purposes of this chapter. Nonetheless, any naturalistic attempt will grant the fact that there are certain associated concepts that we are not allowed to gerrymander, such as causation. These considerations have a direct bearing on Stich's argument.

There are some basic requirements that a 'causal-historical' theory of reference—which Stich endorses—is not allowed to violate. Terms or predicates of a language cannot refer to objects or extensions in the world unless there is an appropriate *causal relation* between the referential expressions in question and the referents they allegedly pick out. This brings two further constraints which will suffice to drive my point home: On the one hand, any putative theory of reference must be able to engage in predictions of linguistic behaviour. On the other hand, the causal relation between the terms employed and the objects they pick out must facilitate non-linguistic dealings with the objects in question. But how can Stich confidently claim that there are theories of reference which meet these constraints, *and*, at the same time, differ in truth value assigments with respect to REFERENCE?

Once the Quinean alternative has been discarded, I cannot think of other strategies that meet this *desiderata.* Take, for instance, Putnam's permutation argument. Putnam exploits the notion of an arbitrary permutation—i.e., an arbitrary

one-to-one mapping of every object in the universe of discourse onto another. In this way, we may obtain *any* arbitrary word-world mapping by making compensatory adjustments to the extensions of predicates when assigning referents to terms.[16] It seems then that Stich has a path to exploit. If I can make any radical rearrangement in the referential relations under consideration, we may find out to our surprise that even though '____ is a belief' fails to refer under REFERENCE, it does *refer** to something under REFERENCE*. Unfortunately, any *radical* rearrangement would miss the causal link between terms and their referents, and considering the above constraints, we would lose any ability to predict linguistic behaviour, and to amend our own cognitive attitudes (linguistic as well as non-linguistic) by using others as a source of information. Hence, it is my contention, Stich's only way out is to gerrymander the notion of causation.[17] By changing the notion of causality in any bizarre way in the Home language—i.e., the metalanguage—, he may stick to his argument. I ignore how such a strategy might actually run; however we need not worry since, as we saw, the overall discussion of

---

[16] The reader not familiar may consult Putnam (1981) chapter 2; and pp. 217-ff. for a formal proof of the argument.

[17] Note that the only way for Stich to meet the above *desiderata* would be by generating 'less radical' alternative schemes of reference—less radical in the sense of trying to preserve the causal links between terms and their referents in the world. But, how 'less radical' can Stich go? Obviously, he would need to produce schemes of reference which earn their keep empirically—i.e., that remain empirically adequate with respect to the standard one (REFERENCE). But, to the best of my knowledge (although see 6.6 below), that can only be accomplished by endorsing a Quinean framework, in which case Stich's argument wouldn't go through for the reasons offered earlier.

Eliminativism pressuposes a common starting point between eliminativists and anti-eliminativists as far as the notion of causation is to be fixed in physics.[18]

In conclusion, if eliminativism is true—premise that Stich grants—, then ELIMINATIVISM*…i…* must also be true, at least for the Quinean alternative ways of generating theories of reference considered above.[19] This outcome holds, I conjecture, unless Stich is willing to give up constraints that govern the construction of our semantic apparatus, as well as, the priviledged status of physicalism, in which case the price we would be paying to refute eliminativism would be far too high.

---

[18] To illustrate the point, it might help to look at a case where there is wide agreement. Take 'phlogiston'. According to Stich's line of argument, the fact that 'phlogiston' fails to refer would be uninteresting. The reason is that, pressumably, there is a different theory of reference according to which 'phlogiston' does refer to *something*. The question for Stich is thus: "But what could that 'something' possibly be?". I fail to find an answer to this question that conforms to our scientific—physical—standards. In fairness to Stich it must be noted that precisely this sort of considerations have made him change his views dramatically on this subject (see chapter 7 below).

[19] If put under preassure, I would be ready to concede that my conclusion is far more modest than the one that the eliminativist should set for herself. Ideally, the eliminativist would like to conclude that ELIMINATIVISM*…i…* must be true for *any* alternative theory of reference Stich may consider.

## 6.6 *Conclusion*

Stich (1990; 1991) argued that (i) the thesis of eliminativist materialism, according to which propositional attitudes don't exist, is neither true nor false, and that (ii) even in the case it were true, that would be philosophically uninteresting. To support (i) and (ii) Stich relied on two premises: (a) that the job of a theory of reference is to make explicit the tacit theory of reference which underlies our intuitions about the notion of reference itself; and (b) that such a notion of reference is a highly idiosyncratic one. In this chapter I tried to show that even if we agreed with premises (a) and (b), Stich's arguments are still doomed.

Before closing this chapter, however, let me expand briefly on an aforementioned caveat with regard to Stich's second argument, and premise (b) above—see fn. 19. As I acknowledged earlier, the ideal eliminativist conclusion according to which if eliminativism is true, then ELIMINATIVISM*…i…* must be true, for any alternative theory of reference Stich may consider, is far too strong. Or, better said, it is too strong to be supported by the arguments I've offered in section 6.5. Rather, what I've tried to show is that Quinean alternatives won't do for Stich, and that Putnam's unfettered permutations won't do either. However, I haven't shown that there is no non-Quinean/Putnamian permutation available which is constrained by holding causation, and the rest of our background theoretical apparatus—notions of reference, explanation, etc.—fixed. On the other hand, it

---

Nevertheless, I believe the eliminativist can live with the more modest results achieved in this section (see section 6.6 below).

must be stressed that Stich has not shown that there is such an option available. Hence, a fairer way to read the results of section 6.5 would be as a stand off between the eliminativist and Stich. This, nonetheless, should not be interpreted as a partial defeat for the eliminativist. Stich himself has abandoned the views we've been concerned with in this chapter. As I mentioned in section 6.2, Stich (1990; 1991) interprets the thesis of eliminativism as the claim that the theoretical terms of folk psychology fail to refer. However, according to Stich's (1996) latest (?!) view, the theory of reference just isn't the place to go to when trying to settle ontological disputes. Stich's latest twist in the eliminativist plot makes of him a 'social constructivist' or, as he prefers, a Quinean pragmatist—see Stich (1996), pp.55-9, p. 72, and chapter 7 (section 7.5) below. If Stich is right, and the theory of reference cannot shed any light upon the eliminativist/anti-eliminativst debate, then whether my results in section 6.5 are strong enough, or not, becomes a secondary issue. The purpose of this chapter has been simply to show that even if semantic considerations of the sort Stich considered threw some light over disputes on ontology, we still couldn't turn a blind eye to eliminativism. Although I agree with Stich that ontological disputes are not to be settled by looking at our semantic commitments, I disagree with the conclusions he arrives at. In the next chapter I shall address these considerations in more detail, and set more ambitious limits for the friend of eliminativism. The general objective of the next chapter will be to produce a connectionist defence of the thesis of eliminative materialism.

# 7

# *CONNECTIONISM AND THE TWILIGHT OF PROPOSITIONAL CONTENT*

## 7.1 *Introduction*

In chapters 4, and 5 we saw the implications that certain key features of connectionist networks had for Quine's Thesis of the Inscrutability of Reference. However, the philosophical implications of connectionism, in my opinion, run deeper, having a direct bearing upon the Theory of Mental Representation. The discussion in chapter 4 (especially sections 4.7, and 4.8) highlighted a crucial issue with regard to the philosophy of mind. Namely, the fact that a connectionist model of cognition fails to endorse the 'computational theory of the mind'. That is, cognitive activity in the connectionist guise does *not* consist of formal operations performed on internal representations according to syntactic rules (see 4.8 above).

Connectionist processing does not exhibit the compositional character of classical representations. Hidden representations in fully-superposed neural networks are not representations of *propositions*. Granting this setting, the friend of the thesis of eliminative materialism has found an ally in connectionist theory. Some philosophers have argued that the above considerations bring support to the elimination of folk psychological posits: if connectionism is true, then folk psychology must be wrong, and the propositional attitudes should be eliminated from our ontology. In this chapter I propose to examine these issues. The purpose is to produce a connectionist defence of the elimination of the mental.[1]

Before getting started, let me briefly outline the programme of this chapter. In section 7.2 I shall introduce a conditional argument for the elimination of the posits of folk psychology put forward by William Ramsey, Stephen Stich, and Joseph Garon (henceforth abbreviated RS&G). In section 7.3 I shall consider an objection to RS&G's eliminativist argument raised by Clark. I shall then review a counter that Stephen Stich and Ted Warfield produce on behalf of the eliminativist. The discussion in chapter 5 on 'state space semantics and conceptual similarity' will be used to show that Clark's argument is not threatened by Stich and Warfield's considerations. Then, in section 7.4, I shall offer a different line of argument to counter to Clark. A line that focuses on the notion of causal efficacy. I hope to show that RS&G's eliminativist argument is correct. Conclusions, and review of two

---

[1] It must be noted that the forthcoming discussion assumes a physicalistic spirit towards the naturalization of content. Worries about *qualia* (e.g., Jackson, 1982), and *view-from-nowhere* arguments (e.g., Nagel, 1989) are tangential to the present enterprise.

other caveats concerning the outcome of the eliminativist/antieliminativist debate will follow in section 7.5.

## 7.2 *Propositional Modularity and Fully-Superposed Neural Networks*

In their 1990 seminal paper Ramsey, Stich and Garon offered a conditional argument for the elimination of the posits of folk psychology—beliefs, desires, etc. In a nutshell, RS&G's conditional argument runs as follows: Folk psychology, insofar as it individuates mental states in terms of their *propositional* content, is committed to the thesis of propositional modularity.[2] The thesis of propositional modularity makes three distinctive claims. Propositional attitudes are (i) functionally discrete, (ii) semantically interpretable, and (iii) causally efficacious.[3] But, RS&G contend, *if* fully-superposed connectionist models of cognition turn out to be correct, *then* there are no such entities with such properties. In particular RS&G consider the way in which internal states of certain connectionist models of

---

[2] The reader should not confuse the thesis of propositional modularity with Fodor's (1983) notion of *modularity*—i.e., accounting for high-level cognitive tasks by building complex architectures out of simpler interacting modules which are orchestrated together so as to deliver highly complex behavioural outputs (see also Minsky, 1985, and Shallice, 1988). Although the approach to cognition advocated in my dissertation (see chapter 4 above) implicitly rejects Fodor's modularity of thought, I shall not try to spell out the divergencies here.

[3] This reading is not forced by the eliminativists themselves, but rather encouraged by defenders of folk psychology's posits—the reader may care to consult, for example, Fodor, 1987, p. 10. See also Stich (1983) for an earlier elaboration of the three tenets of propositional modularity.

memory interact amongst themselves, and claim that this is inconsistent with the interactions of propositional attitude states, as described by folk psychology (see below). The incompatibility between the propositional attitudes' above features—(i), (ii), and (iii)—and some connectionist networks suffices to show, in RS&G's view, that the posits of folk psychology ought to be eliminated from our ontology.

A number of powerful attacks have been launched to cancel RS&G's argument. Addressing all of them wouldn't be realistic for the purposes of this chapter. However, I shall try to counter to one criticism which focuses on a particular aspect of RS&G's argument. The criticism in question is due to Clark (1989/90).[4] First, let me elaborate on RS&G's argument, for it will be crucial to appraise it in some detail before turning to the reactions it has prompted.

The claim that propositional attitudes are committed to the thesis of propositional modularity amounts to saying that: (i) Propositional attitudes are *functionally discrete* to the extent that they can be *individually* lost or acquired, without disturbing other propositional attitudes that an agent might endorse at the given point in time. Think, for example, of cases of memory loss. You may forget *that p*, without losing any other of your current memories. The thesis of propositional modularity holds that the subject believes (dispositionally) the obvious consequences of those propositional attitudes that get tokened in the 'belief box' (if they are mutually consistent). Functional discreteness suggests a cognitive

[4] For other important criticisms which I shall obviate for present purposes, see Stich and Warfield (1995), Smolensky (1995), and Stich (1996). In Calvo Garzón (in preparation c) I address Stich and Warfield's, and Stich's attacks.

architecture designed such that propositional attitudes are encoded in *separate* regions, allowing thus for no domino effect when substracting or adding individual propositional attitudes. In short, beliefs tokened in the belief box which are logically independent of each other are also functionally independent. [5] (ii) Propositional attitudes are *semantically interpretable* to the extent that their content is truth-evaluable and *projectible*—see Goodman (1965). We say that the predicate '___ believes that p' is projectible insofar as its semantic properties—i.e., the belief *that p*—bring about generalizations such as:

> (1) When people believe *that if p then q*, and come to believe *that p*, they will typically come to believe *that q*. (cf. RS&G, 1990, p. 316)

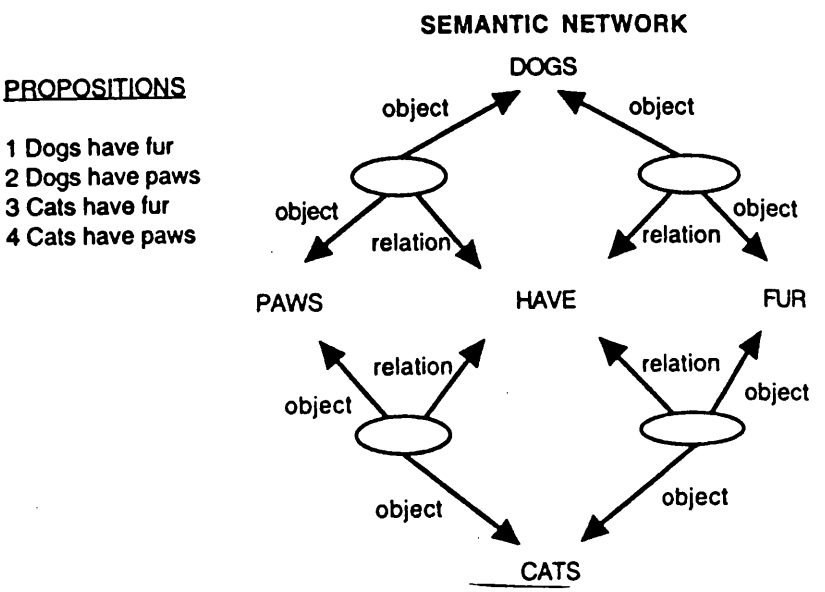
That is, predicates such as '___ believes that p' are projectible insofar as they can figure in causal laws, supporting nomological generalizations. In addition, folk psychology crucially identifies the properties expressed by propositional attitudes' predicates with *natural kinds*. And finally, (iii) propositional attitudes are *causally efficacious* to the extent that they can play a causal role in the production of other propositional attitudes, and ultimately in the production of output behaviour—crucially (see below), distinct propositional attitudes have distinct causal roles.

---

[5] RS&G are careful not to fall prey to *holistic* considerations—cf., for example, Davidson, 1980. Plausibly, under certain conditions, losing or acquiring the belief *that p* may trigger off the loss or acquisition of other semantically related beliefs. However, functional discreteness is not committed to denying this. It is sufficient for RS&G's purposes to note that, at a given time, individual losses or acquisitions can, and indeed, do happen—see RS&G, 1990, p. 316.

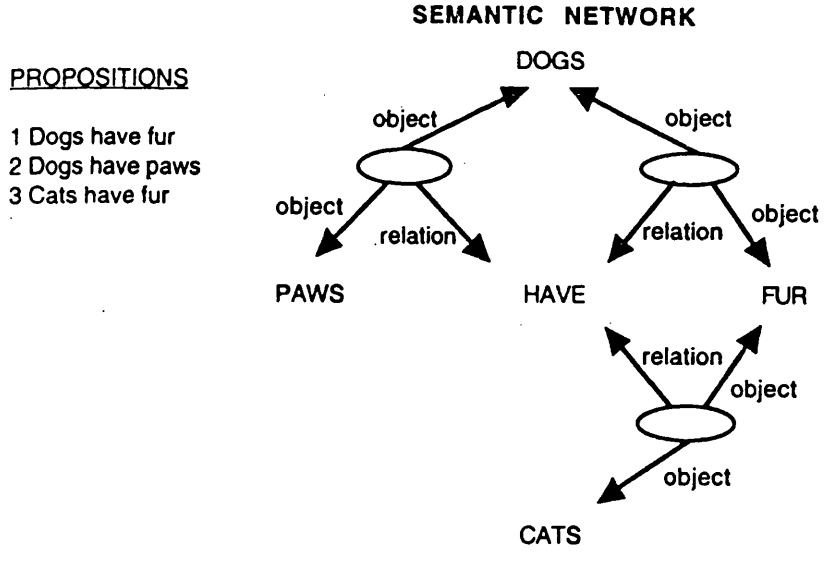
Now, since propositional attitudes are functionally discrete, we can talk of their causal powers in terms of *discrete causal efficacy*. Folk psychology allows us to understand an agent's action as *caused* by *one*, rather than another, propositional attitude. Fully-superposed neural networks (see chapter 4; and below), RS&G urge, are incompatible with (i), (ii), and (iii) above. Before we take a look at the reasons for this incompatibility, it will be useful to introduce a classical model of cognition, that conforms to the thesis of propositional modularity, to be clear about what's at stake, and what the outcome in the debate should be.

There is a number of models of human cognition in the classical literature which take for granted the factuality of the thesis of propositional modularity. For purposes of illustration, RS&G consider Collins and Quillian's (1972) 'semantic



[Fig. 7.1]: A semantic network representation of memory in the style of Collins and Quillian (1972). (From RS&G, 1995, p. 318)

network representation of memory'.[6] Figure 7.1 gives an instance of a section of Collins and Quillian's model. In their model, propositions are represented by nodes along with their labeled links to various concepts. Propositions being stored in memory form a network of functionally discrete and semantically interpretable states. RS&G highlight three key features of Collins and Quillian's model. On the




[Fig. 7.2]: Semantic netwotk with one proposition removed. (*Ibid.*, p. 319).

one hand, as figure 7.2 above illustrates, individual propositions can be added or removed from memory discretely—i.e., without causing readjustments elsewhere in

[6] For another classical illustration, see Newell ans Simon (1972). See also the references in RS&G (1990, fn. 7).

the network. Secondly, predicates are treated as projectible. That is, they are treated as the sort of predicates whose semantic properties allow us to include them in the statements of law-like regularities. In this sense, these semantic properties are taken to constitute genuine natural kinds. And finally, given a certain task such as memory retrieval, we may monitor the network's performance by tracking down the individual propositions that become activated. Some propositions may never get activated during the whole process.

RS&G contend that a certain class of connectionist networks are incompatible with the sort of features that classical models of cognition, such as Collins and Quillian's above semantic model of memory, exploit. We may distinguish two different, although closely related, arguments in RS&G's defence of eliminativism which highlight this incompatibility: An argument regarding *superpositional storage* and discrete causal efficacy, and an argument concerning *natural kinds*.[7] Let us take them in turn.

*The Superpositional Storage/Discrete Causal Efficacy Argument* RS&G employ a connectionist model of memory which is incompatible with two of the three features of propositional modularity. The model—call it Net A—is a three-layered feedforward network consisting of 16 input units, 4 hidden units and one output unit (see RS&G, 1990, p. 325). The task is to answer affirmatively or negatively to each of the first 16 propositions in table 7.1 below, being fed to the network at the input layer.

---

[7] The taxonomy and the dubbing are due to Clark, 1989/90, p. 343.

| *Proposition* | *Input* | |
|---|---|---|
| 1 Dogs have fur | 11000011 00001111 | 1 true |
| 2 Dogs have paws | 11000011 00110011 | 1 true |
| 3 Dogs have fleas | 11000011 00111111 | 1 true |
| 4 Dogs have legs | 11000011 00111100 | 1 true |
| 5 Cats have fur | 11001100 00001111 | 1 true |
| 6 Cats have paws | 11001100 00110011 | 1 true |
| 7 Cats have fleas | 11001100 00111111 | 1 true |
| 8 Fish have scales | 11110000 00110000 | 1 true |
| 9 Fish have fins | 11110000 00001100 | 1 true |
| 10 Fish have gills | 11110000 00000011 | 1 true |
| 11 Cats have gills | 11001100 00000011 | 0 false |
| 12 Fish have legs | 11110000 00111100 | 0 false |
| 13 Fish have fleas | 11110000 00111111 | 0 false |
| 14 Dogs have scales | 11000011 00110000 | 0 false |
| 15 Dogs have fins | 11000011 00001100 | 0 false |
| 16 Cats have fins | 11001100 00001100 | 0 false |
| *Added proposition* | | |
| 17 Fish have eggs | 11110000 11001000 | 1 true |

[Table 7.1]: Propositions Network A and Network B. (*Ibid.*, p. 324)

The output consists of a single unit which is read as 'Yes' if it's on, or as 'No' if it's off. Net A learns to perform this task by backpropagation—see chapter 4, section 4.4 above. If fed, for example, with the coded sentence 'Dogs have fur', activations will spread forward in such a way as to produce a 'Yes' at the output level. In this way, the network shows proficiency in the same task performed by classical cognitive models of memory such as Collins and Quillian's model (see figure 7.1).

This simple feedforward network has two key features which suffice to drive RS&G's point home. The representations that Net A develops in hidden space are *fully-distributed* (see chapter 4, section 4.6), and furthermore, information is stored

in a *superpositional* fashion (see section 4.9). To remind the reader, in fully-distributed neural networks, individual units do not represent particular items. Each sentence being fed to Net A is encoded in hidden state space as a 4-dimensional vector. The network represents sentences as fully-distributed set of values, such that individual hidden units defy semantic interpretation. On the other hand, Net A learns its task by adjusting a single set of weights to produce the appropriate input/output correlation for any of the 16 sentences. All the *knowledge* the network acquires is stored superpositionally in one single set of weights (see chapter 4, section 4.9 above).[8] In short, the key point to bear in mind is that individual units and connection weights embody subtler—subsymbolic—information than the one being represented and processed symbolically.[9]

Net A, insofar as it employs fully-distributed representations and superpositional storage techniques, is incompatible with the thesis of propositional modularity.[10] In particular, it is incompatible with thesis (iii): that logically independent propositional attitudes have distinct causal roles. Since information is

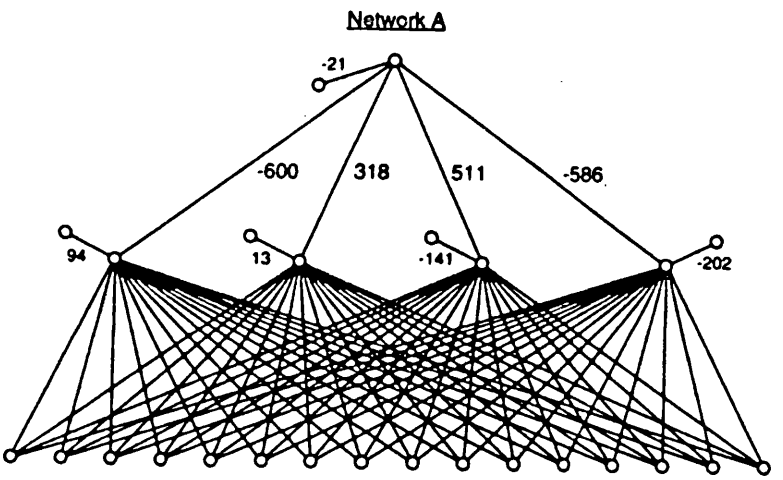
---

[8] I shall talk quite freely of connectionist nets as *knowing* or *believing* propositions. Though, as my argument unfolds (see also chapters 4, and 5 above), it should be clear that these licences are no more than a *façon de parler*.

[9] For a lucid elaboration of the subsymbolic/symbolic dichotomy, see Smolensky (1988).

[10] The careful reader will have realized that the incompatibility won't arise unless the above connectionist model is interpreted as a rival *cognitive* model to classical systems, and not merely as a neural implementation of them. The discussion in chapter 4 (section 4.8) reveals that in fact this is the case. RS&G are careful to make of this crucial property one of their pivotal premises in their conditional argument—see RS&G (1990), pp. 320-ff.

stored in a fully-distributed superpositional fashion, it makes no sense to talk of the *distinct causal efficacious role* played by the representation of a *particular* sentence. Notice that each hidden unit and each weight encode information about every single sentence that has been presented at the input level. Figure 7.3 shows the network's fully-superposed solution to the problem.

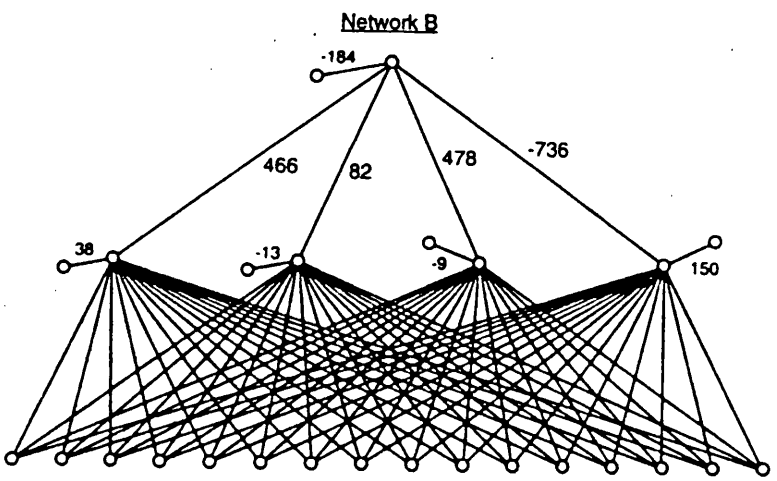



[Fig. 7.3]: Weights and biases in network with 16 propositions. (*Ibid.*, p. 326)

Therefore, it makes no sense to maintain that the net's belief that dogs have fur, encoded by a sentence, say, 'Dogs have fur' is causally responsible—in isolation—for the output 'Yes'. If the belief *that dogs have fur* is meant to play a causal role in the production of a particular output, so does the belief *that cats have paws*, and, for that matter, any other propositional state the net has stored in hidden space. The content of each different proposition is determined by the superposition of all the available representational resources; resources which are a function of the whole

range of input/output patterns that the network has been trained on (see table 7.1). Thus, as figure 7.3 illustrates, the resources the network employs to represent the belief *that dogs have fur* are the same as those required to represent, say, the belief *that cats have paws*. The conclusion RS&G draw is that the connectionist model considered is incompatible with features (i) and (iii) of propositional modularity. That is, propositional attitudes lack discrete causal efficacy.[11]

*The Natural Kinds Argument* To illustrate the alleged incompatibility more vividly, RS&G considered a different network (call it Net B) which is trained, again




[Fig. 7.4]: Weights and biases in network with 17 propositions. (*Ibid.* p. 328)

[11] It is noteworthy that connectionist models need not lack semantic interpretability—feature (ii) of the thesis of propositional modularity. Fully distributed representations are semantically interpretable, though *not* in a *localist* way. The friend of connectionism is willing to interpret the subsymbolic states of a neural network as genuinely representational. The quarrel is rather with the *propositional* approach to semantic interpretability advocated by folk psychologists—although see chapter 4, fn. 5 above, and chapter 5, section 5.7.

by backpropagation, on the same 16 sentences Net A was trained on, plus one more sentence—see table 7.1 above. Since information is fully superposed across the hidden set of units and weights, we won't be able to find a functionally discrete element in Net B which represents the 17th sentence. And so, no element, which can be added or subtracted from the network, that does not disturb other elements of the network. Figure 7.4 above shows the overall solution Net B finds in order to accommodate the 17th proposition.

These considerations support the argument against discrete causal efficacy just reviewed. However, by contrasting Net A with Net B, the case against the propositional modularity of folk psychology can be strengthened. As I mentioned earlier, folk psychology claims that propositional attitudes' predicates are projectible. This allows us to talk of beliefs as constituting single kinds. Nevertheless, since Net A and Net B have no states that can be characterized as functionally discrete, we can say that the representations they encode lack any commonality, at the weights-and-units level, which is projectible (compare fig. 7.4 with fig. 7.3 above). Information is highly distributed, and connection weights embody information relevant to many propositions. We cannot thus identify subregions in Net A and Net B that represent one and the same proposition. The conclusion that RS&G find appealing is that *connectionist beliefs*—whatever they

happen to be—do not constitute a single kind. Connectionist beliefs are rather 'chaotically disjunctive sets'—see RS&G, 1990, p. 329.[12]

RS&G draw a twofold conclusion from the above two arguments: (a) If connectionist models *in the line of* Net A and Net B turn out to be correct,[13] then the requirement of propositional modularity cannot be fulfilled and, therefore, folk psychology is false. And (b), if folk psychology is false, then its posits—i.e., the propositional attitudes—should be eliminated from our scientifically favoured ontology. According to RS&G the fact that the networks under consideration lack two of the three constitutive features of propositional modularity shows that the

---

[12] For the reader less familiar with connectionist theory, the following passage from Clark (1993) nicely echoes, at an intuitive level of understanding, the essence of RS&G's argument: "Imagine the following two ways of storing sentences. In the first way, you keep a discrete token of each sentence on a slip of paper in a drawer. It is then easy to see how to use the tokens one at a time. In the second way, you token each sentence as a pot of colored ink. You then take a vat of water and throw in all the pots. It is now not easy to see how to use the colors separately; worse still, the resultant overall color will vary according to the global set of pots of ink put in. The commonality among various vats which token the same sentence is now lost to view. The question then is: How could a vat-and-inks (read superpositional connectionist) style of storage be compatible with the assumption of propositional modularity?" (Clark, 1993, p. 195).

[13] We already know that Net A and Net B have little or no biological plausibility. There are a number of neurobiological constraints against simple feedforward architectures, and against learning techniques such as backpropagation (see chapter 4, section 4.5). Nonetheless, RS&G's argument can be framed in terms of biologically plausible neural models which are functionally similar to full distribution and superpositional storage—whatever the correct architecture and the learning techniques come to be.

propositional attitudes are *fundamentally* unlike the entities posited by connectionist theorists. We are thus justified, RS&G contend, in drawing the eliminative conclusion, rather than some form of reductionism.[14]

The careful reader can see that RS&G's eliminativist argument goes hand in hand with the connectionist results achieved in Part I of my dissertation (in particular, chapters 4, and 5 above). In Part I, nonetheless, the scope was narrower, making use of connectionist theory to call into question the alleged scrutability that the semantic notion of reference enjoys—all according to the foes of Quine. The target in this chapter is wider, highlighting further philosophical implications of connectionism for the theory of mental representation; implications that go well beyond Quine's Inscrutability Thesis (see 7.5 below). In the next two sections, I shall consider an objection raised by Clark that calls into question a crucial aspect of RS&G's eliminativist argument.

## 7.3 *Higher Levels of Description, NETtalkers, and NETtalk-structures*

Clark (1989; 1989/90) disagrees with the first part of RS&G's argument—(a) above—, and contends that the networks that RS&G deploy to illustrate the alleged

[14] It must be stressed that RS&G don't offer any formal criterion to back up the second more radical part of their argument. That is, a criterion that lets us distinguish potential cases of elimination from cases of mere redution. The contention that the eliminativist conclusion is guaranteed because fully-superposed neural networks lack two of the three core properties characteristic of the thesis of propositional modularity is in the opinion of many commentators the weakest point of RS&G's argument (see section 7.5 below).

incompatibility with the thesis of propositional modularity are in fact consistent with the three features characteristic of propositional modularity—(i), (ii), and (iii) above. To appraise Clark's contention we need to drop the idea that the only kind of description available to the cognitive scientist is the one at the level of the weights and units.[15] Beyond this low level of description of connectionist systems, Clark (1989) considers the status of various other higher-level descriptions. Among these we have the spatial reorganization of hidden space created by performing a statistical cluster analysis, and the symbolic—conceptual-level—descriptions of folk psychology (see Clark, 1989, pp. 188-ff. for a taxonomy of the several low, as well as high, levels of description of connectionist networks). In a nutshell, Clark's misgivings with RS&G's eliminativist argument stem from the fact that *post hoc* statistical techniques such as cluster analysis may reveal that connectionist processing can be subject to a symbolic treatment.[16] For purposes of illustration,

---

[15] The reader may care to consult Smolensky (1988; sections 1 and 2) for a formal appraisal of the low level description of connectionist networks (numerical specification of connection weights, and subsymbolic interpretation of hidden units). See also chapter 4 above (sections 4.4-4.7).

[16] In a stronger reading suggested by Clark, a symbolic evaluation of connectionist systems is not meant to serve simply as a useful 'approximation' of the underlying weight-and-units mechanisms, but rather as a virtual high-level correlate of those mechanisms. Connectionist networks or cognizers *themselves*, Clark speculates, create symbolic representations for the purposes of reinterpreting clusters of features of the sort developed at the subsymbolic level. A somewhat straightforward way to accomplish this may be by building a dual, connectionist-cum-classical, neural network (see for example Miikkulainen's DISCERN network—a distributed neural network that processes simple stereotypical narratives—see Miikkulainen, 1993). Clark himself acknowledges the speculative character of his remarks and does not elaborate on the argument further. I shall not attempt to flesh

Clark (1989/90) considers NETtalk—a very well-known neurosimulation in the connectionist literature developed by Sejnowski and Rosenberg (1986). For the reader not familiar, let me briefly describe NETtalk, for its appraisal will be crucial for the purposes of this and the next section.

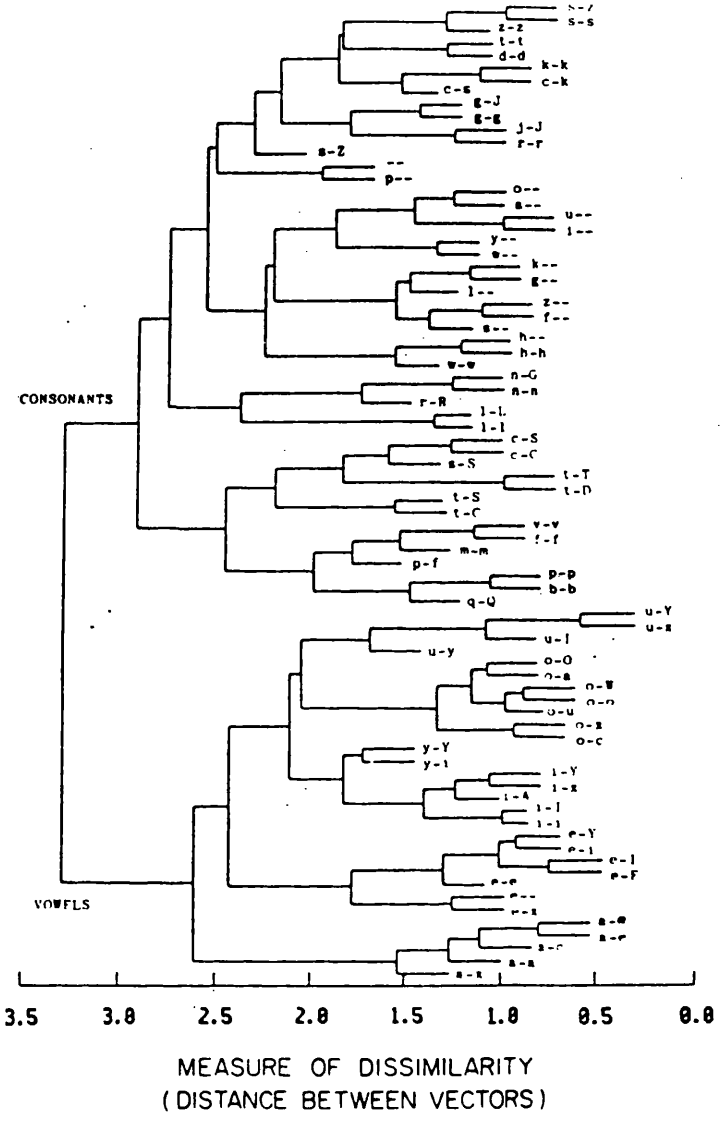
NETtalk is a relatively large network trained to perform text-to-speech transformations. The network has a simple feedforward architecture, and contains 145 input units, 60 hidden units, and 27 output ones—see Sejnowski and Rosenberg (1986) for the details. Being fed with a string of letters, NETtalk learns by backpropagation to yield a coding for each phoneme corresponding to every letter in the text. The text is fed to the network by using a fixed window size for the letter strings. In this way, NETtalk learns to map each letter being presented in the center of the input window onto a given 'phonetic' unit at the output layer (the rest of the window acts as 'context'). The codings are then fed to a speech synthesizer producing the relevant spoken signals. The hidden representations obtained after backpropagation learning are highly distributed. In like vein, the distribution of connection weights presents a homogeneous structure across the network's connections. These fully-superposed representational resources allow NETtalk to lay hold of a number of interesting text/speech regularities. In order to analyze what regularities NETtalk exploits, Rosenberg and Sejnowski developed, as part of their

---

out Clark's line of thought here (the reader may care to consult Clark, 1989, Appendix (section 5) for a quick appraisal of his position). For present purposes I shall focus instead on the less controversial reading according to which higher-level descriptions are simply useful approximations of lower-level mechanisms (see below; see also the discussion in section 4.7 above).

methodology, the statistical technique of cluster analysis. A cluster analysis consists in pairing each hidden pattern of activation with its closest neighbour. An average activation value is then calculated, and the process of pairing neighbours is repeated for the new patterns of activation. Responses to different inputs are accounted for by spatially partitioning the internal representational space (see section 4.5 above). In this way cluster analysis can compare the relation of hidden units to different inputs, and their effects on oncoming outputs. This technique is hierarchically applied, arriving in the end at a final clustering in space where points are located in several specific regions as a function of the similarities shared with other points. In this way, all the vectors whose activation values correspond, for example, to vowels will be seen as points in a vowel-region, and the same goes for the consonant vectors (see figure 7.5 below). The consonant-region, moreover, appears divided in several subregions which include bilabials (p, b), dentals (d, t), etc. Other subregions comprehend, for example, a voiceless, palatal group. In short, cluster analysis reveals that NETtalk navigates successfully the text-to-speech domain by organizing sets of stimuli in terms of the articulatory features characteristic of different phonemes (e.g., voiced, palatal, liquide)—see Sejnowski and Rosenberg (1986), or Rosenberg and Sejnowski (1987), for the details.

According to Clark, cluster analysis furnishes us with a level of analysis which is able to reconcile the connectionist subsymbolic level of description with a straightforward, higher-level, classical approach. NETtalk's hierarchical clustering (see figure 7.5 above) is meant to make manifest the profile of the 'semantic metric' (see 4.5 above) that the network constructs in order to master its task. In short,




[Fig. 7.5]: Hierarchy of partitions on hidden-unit vector space of NETtalk. (From Churchland, 1989, p. 176, after Rosenberg and Sejnowski, 1987)

cluster analysis sets forth the representational configuration developed by NETtalk, described at a more abstract level than the neural one, by finding a hierarchy of partitions in hidden space. By pairing together all those inner states as a function of the (spatial) similitude between the diverse activation patterns, cluster analysis offers a clear example of how neural networks can satisfy a demand for a form of commonality; a commonality which is absent at the individual level of the weights and units—see section 4.7 above.

As the reader may have guessed by now, the bearing of these considerations upon RS&G's eliminativist argument is pretty straightforward. In the first plank of their general argument (the 'superpositional storage/discrete causal efficacy argument'—see 7.2 above), RS&G argued that since the number of active weights prompting any output involves the whole network (i.e., full distribution) and given that each weight participates in the storage of many items of data (i.e., superposition), it is not possible to isolate an individual belief as causing a particular output. However, according to Clark, appealing to the higher level of description of the activation states via cluster analysis shows that fully distributed, superpositional representations are structured enough to be compatible with the requirements of propositional modularity. If the network goes into a hidden unit activation state *within the domains of a subcluster* which we have rightly identified with, say, the symbolic label 'dogs have fur' then, in spite of the superpositional storage of information, Clark contends, we are justified in claiming that the network has arrived at a certain output because at that moment it *believed* that dogs had

fur.[17] On the other hand, the second part of RS&G's general argument (the 'natural kinds argument'—see 7.2 above) highlighted the lack of any genuine kind, at the level of the units and weights, uniting Net A and Net B. But again, Clark insists, this is not a problem. Thanks to cluster analysis we may discover that different networks define a unique class by dividing their respective hidden spaces into significantly similar sub-spaces. We may then assign common kinds to different networks regardless of their units-and-weight's idiosyncrasies. Clark reminds us of a salutary maxim:

> The basic philosophical point here is a very familiar one. Good explanations may demand the grouping together of systems which, at a low enough level of physical description, form a 'chaotically disjunctive set'. (Clark, 1989/90, p. 349)

Summing up, Clark's point is that beyond the connectionist fine-grained level of analysis, there is a higher statistical level of understanding which provides us with the conceptual stability that propositional modularity demands, in which case, the friend of the propositional attitudes has nothing to fear. In this section and 7.4

---

[17] The reader particularly interested in this first argument is encouraged to consult RS&G (1990), and Clark (1989/90). It would require some more work to flesh out fully Clark's reasons for rejecting the 'superpositional storage/discrete causal efficacy argument'. I believe, nevertheless, that the discussion in chapter 4 (especially sections 4.7 and 4.9) bears directly on Clark's response to RS&G's argument. Fleshing out this issue would take us far afield. In what follows, I shall concentrate exclusively on the 'natural kinds' argument (see below) since it is more relevant to my overall purposes.

below, I shall concentrate exclusively on how to rebut to Clark's second attack—i.e., his rejoinder to RS&G's 'natural kinds argument' (see above).

Clark's argument depends crucially upon an *empirical bet*. Namely, that *all* cognitive systems complex enough to count as believers will exhibit certain higher-level commonalities with all other such systems. So, Clark writes:

> The bulk of [my argument] has amounted to an unabashed empirical bet that any system complex enough to count as a believer will reveal (under some *post hoc* analysis) semantically clustered patterns of activation. Such reasonably complex models as we have available (e.g., NETtalk) lend support to this contention. (Clark, 1989/90, p. 352)

In the remainder of this section I shall reflect upon a criticism due to Stephen Stich and Ted Warfield (1995) that precisely calls into question Clark's 'empirical bet'. I shall argue that Stich and Warfield's considerations are wrong, and cannot threaten Clark's anti-eliminativist conclusion. But before that, let me briefly rehearse an interesting line of response that the careful reader may have thought of.

In fairness to the connectionist-eliminativist, it is not a straightforward matter for the anti-eliminativist to make her case, or at least it would require more elaboration than Clark offers. A key issue in the current dialectic relates to the status we ascribe to high-level descriptions of the kind exploited statistically via cluster analysis (see chapter 4, section 4.7). Notoriously, Churchland has argued at length that cluster analysis cannot describe accurately connectionist processing. As we saw in chapter 4, cluster analysis can provide no more than a high-level approximation of a neural network's gross behaviour, failing thus to describe

accurately the dynamical processing undergone by the system. The reason for this failure, as Churchland emphasizes, is that the network itself lacks any information about the clusterings being generated statistically.[18] To put it more dramatically, those clusterings won't take a part in the laws—i.e., learning algorithms (see section 4.4 above)—that exercise control over the behaviour of the network. Thus, Churchland notes:

> the learning algorithm that drives the system to new points in weight space does not care about the relatively global partitions that have been made in activation space. All it cares about are the individual weights and how they relate to apprehended error. The laws of cognitive evolution, therefore, do not operate primarily at the level of the partitions [...] The level of the partitions certainly corresponds more closely to the "conceptual" level [...], but the point is that this seems not to be the most important dynamical level. (Churchland, 1989, p. 25)[19]

---

[18] To keep the record straight, the reader should notice that, although deeply entrenched, the 'anti-approximationist' line of response stressed by Churchland differs from the reasons rehearsed in chapter 4 (section 4.7). There, I argued that Servan-Schreiber *et al.*'s neurosimulations lent support to the view that connectionist processing tends to preserve 'redundant' information; information which is averaged out when performing post hoc *statistical* analyses—the stress lies in the fact that the results obtained are *statistical*, and thus inherently loose processing of idiosyncratic detail. Churchland's aformentioned point, by contrast, emphasizes the *post hoc*, rather than the statistical part of the equation.

[19] A side exegetical issue, although noteworthy, is whether Churchland's above comments are consistent with his latest defence of connectionist semantics (see chapter 5 above). Unlike the Churchland of *A Neurocomputational Perspective* who stresses the cleavage between subsymbolic and symbolic processing, more recently Churchland (1998) makes of higher-level descriptions a

The above considerations, nevertheless, are too general, and do not target specifically the heart of Clark's attack. That is, Clark's aforementioned empirical bet. In what follows I shall examine a rejoinder to Clark's argument due to Stich and Warfield which calls into question Clark's empirical bet. Stich and Warfield—henceforth abbreviated S&W—argue that Clark's empirical bet has no chance of being 'realistically' realized. To remind the reader, Clark is betting that neural networks, or cognitive systems, will share the same cognitive profile, when observed at the appropriate level of description. So, two networks that learn to navigate the same domain will always enjoy a common macrodescription at the level of their respective clustering profiles. In short, their hidden spaces will be partitioned similarily when subject to a hierarchical cluster analysis. S&W reject Clark's empirical bet, and argue that in the case of systems like NETtalk, finding such higher-level commonalities is the exception, rather than the rule. They then conclude that NETtalks—or, extrapolating, believers—form anything but a natural kind. In the remainder of this section I shall rebut to S&W's argument, arguing that

---

pivotal factor in bringing robustness to State Space Semantics—see section 5.4 above. This does not mean that there is an unresolvable inconsistency in Churchland's position. It may well expose an intellectual evolution towards a more moderate position in the debate. However, to the best of my knowledge Churchland has not acknowledged such a change of gears in print. I suspect that a strong will to have it both ways—i.e., deny processing accuracy to statistical analyses while employing them to bring robustness to semantic discourse—may ultimately force him to reconsider his general approach to connectionist semantics. I shall not press on the issue here—see Calvo Garzón (in preparation a; in preparation b).

their considerations fail to deliver the goods to the sympathiser of eliminativism. Let us consider their argument in more detail before submitting it to critical scrutiny.

To make their case, S&W draw a distinction between what they call *NETtalkers* and *NETtalk-structures*. A NETtalker is any neural network that can transform text into speech beyond a certain level of accuracy. That is, behaviourally or functionally speaking, any system that delivers the same outcome as Sejnowski and Rosenberg's (1986) NETtalk. On the other hand, a NETtalk-structure is any system that, apart from delivering the correct results, is architecturally speaking similar to Sejnowski and Rosenberg's NETtalk. That is, it has the same number of units and connections, which are distributed in the same, or similar, way. Now given this distinction, S&W argue that Clark's empirical bet is either a sure looser, or is irrelevant to the debate over the fate of the folk.

Clark's only chance of winning the bet, S&W believe, is by considering NETtalkers with the same NETtalk-structure. In that case, it is obvious that they will display the relevant higher-level commonalities at the level of their clustering profiles.[20] However, if an argument that relies on connectionist networks is to have

---

[20] This is not straightforward, or at least it requires some qualification. NETtalkers with the same NETtalk-structure could still have *disimilar* clustering profiles due to idiosyncracies in their respective training regimes—e.g., different learning rate, momentum, etc. Therefore, besides sharing their architectures, two NETtalkers with similar NETtalk-structures will manifest the relevant high-level commonalities only if they are similar to Sejnowski and Rosenberg's NETtalk in some other aspects. S&W are careful enough in bearing these considerations in mind, although they are not very explicit about it.

any bearing in the debate over the propositional attitudes, we must be ready to compare networks with different NETtalk-structures (unless, of course, we are willing to admit by extension that *all* believers—whatever they happen to be—have the same 'BRAINtalk-structure'). Granted that, S&W point out that we could model many different NETtalkers, none of which has a NETtalk-structure similar to the one of Sejnowski and Rosenberg's NETtalk. Now, given that networks with different internal dimensionality—i.e., different number of hidden units—tend to find different solutions to their overall problem, S&W argue, we will find many different NETtalkers which don't have a clustering profile in common.[21] The bearing of S&W's considerations upon Clark's empirical bet is now obvious. Clark holds that any two cognitive systems complex enough to count as believers will have a number of high-level commonalities; commonalities which highlight a natural kind beyond the personal idiosyncrasies of the systems. Clark, however, cannot characterize the systems that configure this natural kind in terms of their shared architectures. In that case, the fact that they share certain high-level commonalities would be uninteresting. What he needs is a behavioural or functional characterization that gathers all 'architecturally-divergent' believers. Unfortunately,

---

[21] The discussion of Servan-Schreiber *et al.*'s neurosimulations (chapter 4; section 4.7 above) clearly illustrates the reason for this. Networks with higher dimensionality have more representational resources, and can thus process finer-grained detail than networks with scarcer representational resources. Hierarchical clusterings will thus tend to diverge, the bigger the difference in dimensionality between the networks—although see below.

as S&W's distinction between NETtalkers and NETtalk-structures shows, natural kinds don't necessarily emerge in that case.

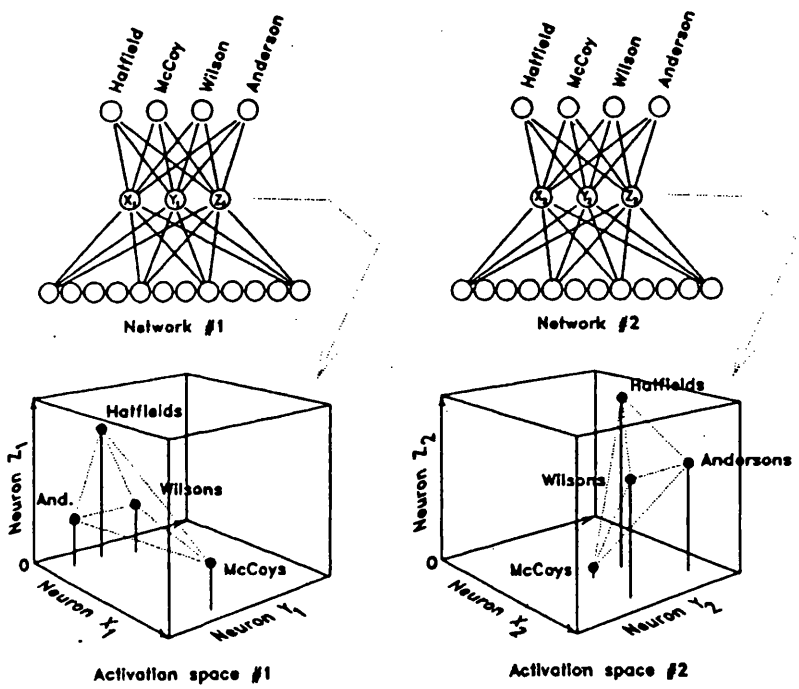
In the remainder of this section I shall argue that S&W's rejoinder is unable to show that Clark's empirical bet has only a minimal chance of winning. Indeed, as we'll see next, Clark's bet may have already been backed empirically (indirectly) by the hand of Laakso and Cottrell's neurocomputational results reviewed in chapter 5. By the time S&W wrote their rebuttal to Clark (1995), it wasn't clear how representational spaces of different dimensionality could be *objectively* compared across networks. Fortunately, the discussion in chapter 5 on 'conceptual similarity and state space semantics' brings new light to our current concerns, highlighting the fact that different partitions of hidden space across networks can still reveal a *common* strategy despite the functional and structural idiosyncracies inherent to connectionist processing.[22] Let me explain.

The picture offered by S&W is far too simple. Two NETtalkers with different NETtalk-structures can have a hierarchical clustering profile in common, despite the fact that hidden space gets partitioned differently in each network. Churchland

---

[22] Many different *post hoc* statistical techniques had already been successfully deployed in order to compare representational spaces across networks. However, the task became increasingly difficult, the more the clustering profiles tended to diverge. Lack of measurement techniques wasn't the problem. The problem was rather lack of *objective* measurement techniques. As Cottrell pointed out when presenting Laakso and Cottrell's neurocomputational results (see chapter 5 above): "When I eyeball the dendograms [i.e., the tree-structured hierarchical clusterings] for two distinct networks, I may say, 'yes, they are fairly close', but that's just my reaction. We need an objective measure of such things" (quoted from Churchland, 1998, p. 18).

(1998) illustrates how this can happen with a simple example.[23] Imagine two simple feedforward networks with identical architectures, which are trained on a classification task on a base data set composed out of 100 photos of each of 100




[Fig. 7.6]: The locations of four prototype points within the hidden-layer activation spaces of two (imaginary) neural networks for recognizing the faces of four different extended families. The four points represent a prototypical Hatfield face, a prototypical McCoy face, a prototypical Wilson face, and a prototypical Anderson face. (Churchland, 1998, p. 9)

[23] Churchland develops the example below for different purposes—answering Fodor and Lepore's challenge to State Space Semantics (see chapter 5 above). However, we may easily transport Churchland's dialectic to the current discussion on NETtalkers and NETtalk-structures.

members of four families. Figure 7.6 above shows the different arrangement in the respective hidden spaces of the two networks of the prototypical faces of the four families.

As figure 7.6 illustrates, the two imaginary networks have arrived at (apparently) different solutions in order to succeed in their common discriminatory task. However, as Churchland notes, we can help ourselves to a higher level of understanding in which the two networks' strategies do converge. In particular, Churchland proposes that we look at the relative position of each prototypical point against the position of the other three prototypical points within each space (cf. chapter 5, section 5.2 above). Indeed, if we pay attention to those *relative* positions we can see that they are identical across the two networks. If we consider the 3-dimensional solids formed by taking the prototypical points as vertexes, we can see that they are identical (see figure 7.6 above). The reader can see that by translating and rotating one solid with respect to the other, we obtain the same irregular tetrahedron.[24] The interesting point for our purposes is that this intuitive higher-level form of similarity—i.e., translating and rotating n-dimensional figures in hyperspaces to check whether they highlight an isomorphic configuration of points—can be extended into an objective numerical criterion of similarity. The following measure offered by Churchland (1998, p. 12) permits us judge how similar or disimilar two solids are. A value of 1 indicates that the two solids are identical.

$$\text{Similarity} = 1 - \text{Average}\left[\frac{|AB - A'B'|}{(AB + A'B')}\right]$$

[24] The reader interested in the fine-grained detail is urged to consult Churchland (1998).

(*AB* and *A'B'* stand for the lengths of the edges between pairs of prototypical points belonging to two different solids—see Churchland, 1998, pp. 11-12 for the details). This numerical measure allows us to compare objectively any pair of solids across hidden spaces.[25]

Churchland's numerical measure of similarity across representational spaces may then throw some light upon our present worries, hopefully showing whether Clark's bet can obtain empirical support or not. The idea is that we can objectively compare different NETtalkers. A value close to 1 will show that two NETtalkers have landed on a similar solution to their shared problem—i.e., that they have partitioned their hidden spaces similarily. The networks, we may expect, will organize similarly sets of stimuli in terms of the articulatory features characteristic of different phonemes so that they can navigate the text-to-speech domain successfully.[26] These considerations contain the gist for a rejoinder to S&W.

---

[25] The careful reader will have noticed that the above numerical measure has a shortcoming. Namely, that it is sensitive to differences in scale. Two shape-identical solids which differ in absolute size will obtain a similarity rating closer to 0 than to 1. Nevertheless, we need not worry about that. The above equation can be repaired by inserting a 'correction factor' (see Churchland, 1998, p. 19 for the details).

[26] Someone may argue that the only way to back empirically Clark's bet would be to obtain a measure of similarity across networks of 1. That is, showing a perfect identity in their clustering profiles. This may be the case in the toy-example that Chuchland makes use of, where a translation and rotation of one figure with respect to the other produces a perfect fit. This however is not realistic. The more complex the networks under consideration are, the more difficult it will be to

Unfortunately, the imaginary simulations Churchland makes use of to illustrate his measure of conceptual similarity are too simple. Indeed they cannot serve to address fully S&W's challenge by themselves. The reason is that Churchland is considering networks with the same architecture. S&W's point was precisely to compare NETtalkers with *different* NETtalk-*structures* (see above). So, even though different NETtalkers may obtain a value close to 1, using the above measure of similarity, it is only so, S&W would contend, because of having a similar NETtalk-structure.

This, however, should not cause any concern. Bearing in mind the results of chapter 5 above, we may go beyond Churchland's imaginary neurosimulations, and especulate about what might happen by looking at real simulations along the lines of the ones developed by Laakso and Cottrell. To remind the reader, Laakso and Cottrell—see chapter 5, section 5.3—ran an experiment on a colour-categorization task employing networks with *different internal dimensionality*, as well as different input codings. The networks employed had between 1 and 10 hidden units. Once the networks mastered the categorization task, certain mathematical measurements were computed,[27] and the correlations obtained were very high, independently of

---

obtain a perfect fit. However, S&W only call into question the alleged similarity (not identity) of NETtalkers with different NETtalk-structure. We can thus relax our demands, and hope for a value close 'enough' to 1.

[27] The numerical measurement deployed by Laakso and Cottrell is related to Chuchland's above measure, although computationally more demanding—see Churchland (1998) and Laakso and Cottrell (2000) for the detail.

the number of hidden units employed by the networks. From these results, Laakso and Cottrell concluded:

> Our measure is a robust criterion of content similarity, of just the sort that Fodor and Lepore demanded in their critique of Churchland. It can be used to measure similarity of internal representations regardless of how inputs are encoded, and regardless of number of hidden units. Furthermore, we have used our measure of state-space similarity to demonstrate empirically that different individuals, even individuals with different "sensory organs" and different numbers of neurons, may represent the world in similar ways. (Laakso and Cottrell, 1998, pp. 595-6)

The reader can see that Laakso and Cottrell's results tell against S&W's above argument. The upshot of Laakso and Cottrell's neurosimulation for our current concerns is that different dimensionality, architecture or encoding bring no trouble, insofar as correlated distances between points in the respective spaces are preserved. If we agree with Churchland's above contention—namely that fit of prototypical trajectories via rotations, translations, etc. provides us with a connectionist notion of conceptual similarity—we have a straightforward link to the discussion on NETtalkers and NETtalk-structures. Similarily to Laakso and Cottrell's array of architecturally different networks, NETtalkers with different NETtalk-structures will presumably have a robust tendency to settle into the same abstract solution with regard to how they structure the partitions within their activation spaces. So long as the relevant information is somehow implicit in whatever sensor-input schemes happen to be employed, and so long as the training

procedures impose the same requirements on recognitional performance, then diverse nets can settle into almost identical abstract organizations.

In conclusion, extrapolating from Laakso and Cottrell's results, we may expect, *contra* S&W, NETtalkers with different NETtalk-structures to define a unique class by dividing their respective hidden spaces into significantly similar sub-spaces. We may then assign genuine kinds to different NETtalkers regardless of their architectural idiosyncrasies. In fairness to S&W, it must be stressed that Laakso and Cottrell's neurocomputational results may not be applicable to the case of NETtalkers with different NETtalk-structures. Laakso and Cottrell tested their results with networks whose hidden spaces ranged between 1 and 10 dimensions (see section 5.3 above). S&W, by contrast, speculate about what would happen with NETtalkers containing, for example, 80, 800, or 8,000 hidden units. Being numbers much bigger, Laakso and Cottrell's results may not be directly applicable. It may be the case that the degree of correlation across networks may decrease the more dimensions we consider—cf. Servan-Schreiber *et al.*'s results in chapter 4. Bearing in mind these considerations, we may read the results of this section as a stand off between Clark, and S&W. This conclusion, I must confess, is more modest than the one I would have liked to draw. Unfortunately, the question is an open empirical one, and more field research ought to be done. This however does not represent a handycap for our present discussion. Granting for argument's sake that Laakso and Cotrell's results may tip the balance in Clark's favour, the ball is in the court of the eliminativist. In the following section I shall offer a rebuttal to Clark's anti-eliminativist argument that does not depend on S&W's considerations. I shall argue

that the key to unlock the eliminativist/anti-eliminativist debate resides elsewhere, the pivotal factor being whether statistical entities are causally inert or not.

## 7.4 *Cluster Analysis, and Causal Efficacy*

As we saw in section 7.3 above, Clark urged *contra* RS&G's 'natural kinds argument' that, thanks to *post hoc* statistical analyses, connectionist networks can be seen as compatible with the thesis of propositional modularity. Thanks to cluster analysis, Clark argued, we may discover that different networks define a *unique class* by dividing their respective hidden spaces into significantly similar sub-spaces. We may then group different networks to the same genuine kind regardless of their units-and-weight's idiosyncrasies. My aim in section 7.3 was to show why the eliminativist cannot rebut Clark's argument by arguing that in the case of systems like NETtalk, finding such higher-level commonalities is the exception, rather than the rule. If the research due to Laakso and Cottrell (see chapter 5, and section 7.3 above) can be extrapolated to architecturally more complex networks than the ones Laakso and Cottrell considered, then NETtalks—or believers—may well conform a natural kind, as Clark's argument requires. This, however, does not mean that Clark has won the battle. In this section I shall offer, on behalf of the eliminativist, a strategy different from S&W's line of argument (see 7.3 above) in order to bypass Clark's anti-eliminativist conclusion. Put bluntly, the problem, as I see it, does not depend on whether statistical analyses can highlight higher-level commonalities being shared by networks which, at a lower level of description,

look radically different. In my view it is likely that commonalities do arise. As a matter of fact, that's the very job that *post hoc* statistical techniques are supposed to do (see chapter 4, section 4.7). The real issue, I contend, boils down to whether or not abstractions of the sort generated statistically are *causally efficacious* in the production of the network's patterns of input/output behaviour. In what follows I shall argue that they are not. This lack of causal efficacy, I believe, tips the balance in favour of RS&G's eliminativist argument. Let me elaborate.

Clark's move (section 7.3 above) is based on a failure to appraise one distinction which has lead many philosophers to miscalculate the putative target of the eliminativist. As I argued in chapter 4 (section 4.7) *post hoc* statistical techniques of the sort deployed in the neuromodeler's methodology are causally inert, thus failing to play any explanatory role as far as the dynamics of connectionist networks is taken as our model of cognition (see section 4.8 above). Abstractions of the sort generated statistically provide us with a good way of understanding what kind of representations neural networks can encode. However, the statistical generation of *localized* descriptions of the network's representations should be interpreted as no more than an external abstraction posited in an attempt to understand what the network is doing. Symbolic understanding is genuinely alien to the network *itself*. Positing static symbolic descriptions of a network's stored knowledge merely reflects the modeler's 'invasive' strategy to appraise the network's highly distributed representational resources. What gets activated, at each step of processing, is a component of the cluster, not the cluster itself. The network works exclusively at the level of the numerous context-dependent and distributed

patterns of activity in hidden space. Uniting some of these hidden states under linguistic labels should not drive us to think that the network actually employs entities posited by folk psychology. The reader can see now why these considerations lend support to the first part of RS&G's eliminativist argument. Namely, to the conclusion that connectionist processing is incompatible with the thesis of propositional modularity. *Contra* Clark, unlike the properties expressed by propositional attitudes' predicates, the properties of fully-superposed neural networks don't constitute natural kinds. Rather, they are simply chaotically disjunctive sets—cf. Clark, 1989/90. The issue, unfortunately for the eliminativist, is not that easy to untangle. On behalf of the anti-eliminativist, Clark offers a rejoinder to the putative lack of causal efficacy of statistical abstractions. Clark argues that denying their causal efficacy may place us in a dangerous position of 'microphysical worship'. In the remainder of this section I shall address Clark's contention.

Clark anticipates the aforementioned point on causation, and denies that the eliminativist can exploit the lack of causal efficacy of statistical posits:

> Someone might, I suppose, worry that *being in a certain cluster* cannot, properly speaking, be a cause. Thus, they might insist that what actually does the causing must always be a *particular* hidden unit activation pattern and hence that, if we have to appeal to clusterings of such patterns to find analogues for semantic items, the semantic items cannot figure in the real causal story. (Clark, 1989/90, p. 350)

Clark then objects:

> But this is surely a dangerous move. For it places the philosophical feet on a slippery slope to physics worship [...] And this is radically revisionary. Chemistry, for example, is generally regarded as a respectable special science, and yet it is concerned to group different physical structures as instances of chemical types and to define causal laws which apply to those types. So, unless the sceptic is willing to give up the causal efficacy of chemical properties too, he or she would be unwise to object to the very *idea* of higher-level constructs figuring in genuine causal claims. (*Ibid.*, p. 350)

In sum, Clark invites us, in view of the potential disaster of falling into physical worshipping, to adopt *post hoc* statistical techniques as a genuinely causal way of appraising the network's computational capacities. *Post hoc* statistical techniques, Clark claims, are a genuine way of making fully-superposed neural networks compatible with the thesis of propositional modularity—a thesis that folk psychology relies on crucially (although see below).

I don't think that Clark's twofolded picture fully reflects the range of possibilities. In the above quote, Clark makes his case by appealing to our intuitions regarding a respectable non-basic science such as Chemistry. However, it is not the case that either *all* higher-level causal claims are genuine, or that we must reject all such claims *altogether*—becoming thus 'microphycical worshippers'. There is, I contend, a crucial disanalogy between the case of Chemistry and the putative causal efficacy of the statistical entities employed in connectionist theorizing. Whereas chemical entities are composed out of microphysical structures (see below)—being thus *real* physical objects in the world—the entities statistically posited to explain connectionist dynamic processing are *abstractions*. Such disanalogy, I believe,

marks the watershed between causally efficacious higher-level properties and causally inert ones. We can thus move the burden of causation safely from a micro to a macro-level in certain cases—as usually happens with special science—avoiding, therefore, falling into microphysical worship.[28] Let me elaborate.

When we are reducing a macrolevel theory to a microlevel one, it is required, I conjecture, that the objects in the ontology of the non-basic theory can decompose into microparts which belong to the ontology of the basic theory. This requirement is met in the case of respectable special sciences such as Chemistry, Biology or Genetics. By contrast, the requirement is not fulfilled in the case of folk psychology. In the former case we employ scientific causal explanations which invoke macrophysical properties, such as solubility, ridigity, gene, etc. The crucial contrast resides in the fact that these macroproperties can be explained by deriving their respective macro-level laws from laws covering the behaviour of their *microconstituents*. So, for instance, a genuine causal explanation can appeal to the behaviour of planets or galaxies, insofar as those macroobjects are built up out of their constituent subatomic particles. Or take Genetics. Genes are composed out of

---

[28] It is noteworthy that the above picture does not exhaust the realm of possibilities in the logical space. Some voices in the philosophy of science—notoriously Nancy Cartwright (forthcoming)—would only attribute causal efficacy to the 'lowest' microphysical properties (whichever they happen to be). I feel pretty sympathetic with this approach, which is ready to bite the bullet, and acknowledge the existence of natural kinds only at the most basic level of physical description. Unfortunately, I currently lack the conceptual apparatus to flesh out this more radical alternative. I shall thus limit myself in this section to exploiting the cleavage between macro—*real*—objects and statistically generated *abstracta*.

various physical microconstituents—namely, particular DNA molecules.[29] This doesn't commit us to denying the causal efficacy of biological properties. Genetics has a characteristic vocabulary—e.g., 'gene', 'phenotype', etc.—in terms of which we can formulate its distinctive causal laws. Genetic causation is nevertheless genuine. We can employ sets of biconditional *bridge laws* which, acting as auxilliary premises, connect the vocabulary of Genetics with the vocabulary of the underlying microphysical theory, where the microphysical causation takes place. Once the laws that cover the behaviour of DNA molecules are conjoined with a number of empirically adequate bridge laws, we can obtain genetical macroexplanations where talk of causation is certified in virtue of the molecular microconstituency of genes. On the other hand, in the case of statistical analyses we lack such license since we don't find the required microconstituents. The statistical groupings obtained do not contain the causally active hidden patterns of activation as constitutive parts. Statistical analyses do group patterns of hidden activation in virtue of their particular causal efficacy—in virtue of which outputs different patterns of activity produce. Nevertheless, the entities obtained statistically do not have the hidden vectors as parts. In this sense, the higher (symbolic) level posits *abstracta*, rather than real physical entities which preserve their microphysical constituents as parts. The former posits, not the latter, I claim, cannot enter into genuine causal chains.

---

[29] To be precise, we would have to pick out the very elemental particles that DNA molecules are composed of. We can stick to the molecular level, bearing this in mind.

In conclusion, rejecting the causal efficacy of statistical properties does not make of me a microphysical worshipper. Plausibly, I am ready to concede to Clark, the causal organization of the world can be taxonomized at many different levels. As long as macroobjects are built up out of their real microconstituents, I claim, we can rest assured that a genuine causal explanation is in place. We can make sense of the causal powers, not only of the scientifically respectable entities basic to the ontology of science—e.g., particles, waves, fields—, but also of *some* higher-level ontologies—e.g., common sense objects, such as DNA molecules, tables and chairs, or planets and galaxies. We can make true causal statements at the higher-level, in virtue of macroobjects being constituted out of microphysical entities with genuine causal powers. Unfortunately for the anti-eliminativist, the belief *that p* is not a whole that decomposes into constitutive parts subject to analysis at a microlevel. The belief *that p* is a statistical unit that defies any such reinterpretation from the macro level into the micro-cognitive level. Hence, I claim, Clark offers no compelling reasons to sustain the view that connectionist models are compatible with the thesis of propositional modularity.

## **7.5** *Conclusion*

In this chapter I have defended RS&G's eliminativist argument according to which fully-superposed neural networks are incompatible with the thesis of propositional modularity. The intentional representations that we find in the networks considered by RS&G (see section 7.2 above) are not representations of propositions. RS&G's

argument was aimed to highlight a form of eliminativism: if fully-superposed neural networks are a plausible model of cognition, then folk psychology cannot be right, and the propositional attitudes ought to be eliminated from our ontology. In sections 7.3, and 7.4 I focused exclusively upon the first part of RS&G's argument, trying to show that connectionist models of cognition are incompatible with folk psychological posits due to the lack of causal efficacy manifested by the propositional attitudes. However, on behalf of the anti-eliminativist, I must admit that more steps would be required in order to bring about the eliminativist conclusion. For one thing, the friend of folk psychology may disagree with the thesis of propositional modularity, and claim that the propositional attitudes need not be causally efficacious. On the other hand, someone may deny that the propositional attitudes must be eliminated, while agreeing with the thesis of propositional modularity, and assuming that folk psychology is wrong. In this closing section I shall briefly address these two issues.

On a line of response to RS&G's eliminativist argument, different to that reviewed in section 7.3, Clark acknowledges that the fact that the kind of analyses that connectionist theory furnishes us with is dissociated from the condition of causal efficacy. That concesion, nonetheless, should not cause any distress to the foe of eliminativism. Clark claims that even if the above connectionist reading were correct, and higher-level constructs remain causally inert, the eliminativist conclusion would still not follow. The reason, put bluntly, is that the explanatory role allegedly played by folk psychology's posits in revealing the coarse-grained nature of cognition does not need to be subject to the condition of causal efficacy.

All that is required, according to Clark, is a notion of causal explanation dissociated from the requirement of causal efficacy which appears to be crucial for the thesis of propositional modularity (see 7.2 above).[30]

Following Jackson and Pettit (1988), Clark distinguishes between *program explanations* and *process explanations*. Broadly speaking, an example of a program explanation is any high-level explanation that, while gathering a range of cases in terms of certain macrofeatures being shared, abstracts away from the actual microfeatures which carry the burden of causation. Following that type of explanation, those common macrofeatures are said to 'causally program' a given pattern of behaviour, without actually being part of the causal explanation of that behaviour. By contrast, an example of a process explanation is an explanation that picks out the microfeatures that are causally efficacious. Clark's claim then is that those explanations that employ the various higher-level constructs of connectionist theory may be fully accurate program explanations, while, on the other hand, fail to be genuine process explanations. In short, we may say that cluster analysis (see 7.3 above) causally programs the network's performance, although it does not play any role as part of the process explanation of the behaviour of the network.

Someone may reply to Clark by exploiting a distinction between two types of program explanations: *Derivative*, although genuine, program explanation, as opposed to *abstract* program explanation. The dichotomy would be aimed to reflect

---

[30] Other authors who would disagree with the core properties of folk psychology, as framed under the thesis of propositional modularity are Horgan and Graham (1991), and Jackson and Pettit (1988; 1990).

the distinction highlighted in section 7.4 between real macrophysical objects in the world, and abstractions generated statistically. In this way we may identify 'genuine' program explanations with those high-level explanations which, although failing to be part of the causal explanation of a given pattern of behaviour, posit entities that can be decomposed into the actual microfeatures that carry the burden of causation. On the other hand, 'abstract' program explanations would be indentified with those high-level explanations that posit entities that defy such decomposition from a macrophysical level to the microphysical one. I am aware that the above distinction ought to be fleshed out in more detail. However, for present purposes we don't need to do so. The reason is that RS&G's eliminativist argument targets *only* a reading of folk psychology that conforms to the thesis of propositional modularity (see section 7.2 above)[31] That is, a reading such that beliefs, desires, and the rest of the propositional attitudes are, among other things, causally efficacious. In this way, the sympathiser of eliminativism can make use of RS&G's argument, ignoring Clark's above remarks. Clark himself acknowledges that once we grant Fodor's approach to the debate, the balance is unavoidably tipped against the anti-eliminativist:

> [Many] defenders of symbolic AI and folk psychology (especially Fodor and Pylyshyn) are effectively shooting themselves in the feet. [The] defences they attempt make the condition of causal efficacy pivotal, and they try to argue for

[31] This reading is encouraged by notorious defenders of folk psychology, such as Fodor (1987). See also Fodor (1998a) for a more recent elaboration that is faithful to his earlier views, reiterating the thesis of propositional modularity.

> neat, in-the-head correlates to symbolic descriptions (see, e.g., Fodor 1987; Fodor and Pylyshyn 1988). This is accepting terms of engagement that surely favor the [eliminativist.] (Clark, 1989, p. 197)

Thus, in the light of these remarks, we may ignore Clark's aforementioned line of argument, and conclude with RS&G that the propositional attitudes are incompatible with connectionist models of cognition, *insofar* as the former are defined in terms of the core properties that the thesis of propositional modularity exploits.

There is yet another hurdle that may prevent RS&G from reaching their eliminativist conclusion. Stich and Warfield (1995)—S&W, abbreviated hereafter—agree with Clark that RS&G's eliminativist argument has a small chance of working. However, unlike Clark, S&W claim that the difficulties for RS&G's argument stem from the second part of their twofold strategy—i.e., RS&G's contention that if folk psychology is mistaken *then* the propositional attitudes ought to be eliminated from our ontology (see section 7.2 above). As S&W point out, it does not follow straightforwardly from the fact that folk psychology is mistaken—assuming the first part of RS&G's argument—that folk posits should be eliminated. Intuitively, we may agree, for example, that although ancient people lacked any knowledge about cosmology, they were still referring to the same heavenly bodies that modern astronomy studies nowadays. Stars do exist despite the fact that ancient stars gazers had extremely erroneous ideas about their constitutive properties. In like vein, the fact that fully-superposed neural networks are incompatible with the thesis of propositional modularity does not necessarily entail that the propositonal

attitudes don't exist. They may play a role as part of a future theory of cognition, although probably they will have to be revised under the light of 'yet-to-untap' developments in neuroscience and connectionist theory. In short, as S&W correctly point out, there is a "significant logical gap" to be filled in order to bring about the eliminativist conclusion.

A way one might try to bridge this gap, S&W argue, is by turning our attention to the theory of reference. In particular, by looking at the way in which the theoretical terms of a theory get fixed according to our favoured theory of reference. S&W consider whether the 'description' theory of reference can fit the bill for the friend of eliminativism.[32] In a nutshell, the description theory of reference claims that the theoretical terms employed by any given theory refer to those entities that satisfy most, if not all, of the descriptions that the theory entails about those entities.[33] The satisfaction of most of these descriptions is taken to provide necessary and sufficient conditions for the existence of the entities being posited. If the theory under consideration is false, such that no causal role is played by the entities posited, then the theoretical terms that the theory makes use of do not refer to anything at all. It seems at first sight that the description theory of reference could furnish us with a way to fill the logical gap that has been missing so far in

---

[32] The reader is urged to visit chapter 6 above for an appraisal of Stich's views on the theory of reference, and its bearing upon ontological disputes. For a different line of argument that exploits the notion of a "constitutive property" in order to fill the aforementioned logical gap, the reader may care to consult S&W (1995), pp. 407-9. For present purposes we may ignore this other line of response which S&W themselves don't find very attractive.

[33] For an early formulation of the description theory of reference see Lewis (1972).

RS&G's conditional argument. According to the description theory of reference, the term 'phlogiston', for example, would refer to nothing since it's been widely acknowledged that XIXth. century phlogiston theory is mistaken. No causal role can be ascribed to the core entities that phlogiston theory posits in the explanation of combustion. Similarly, by assuming the description theory of reference, the friend of the elimination of the mental can make her case. Assuming that folk psychology is wrong—once we grant for the sake of discussion the first part of RS&G's conditional argument (see 7.2 above)—the conclusion to draw is that beliefs, desires, and the rest of the propositional attitudes don't exist, since they play no causal role in the production and explanation of a cognitive agent's behaviour.

S&W, however, favour the 'causal-historical' theory of reference—e.g., Putnam (1975), Kripke (1972)—over the description theory of reference.[34] Put bluntly, after an initial reference-fixing event, reference is transmitted along a causal-historical chain (see chapter 6). A virtue of causal theories of reference is that they cope very well with problems of ignorance and error. That makes them perfect candidates for the foe of eliminativism. Notice that according to the causal theory of reference a person can refer to an object, or kind, despite having wildly mistaken views about the object, or kind, in question. Thus, were we to favour a causal-historical theory of reference, it would make sense to suppose, for instance, that ancient stars gazers and modern astronomers talk about the very same heavenly

[34] S&W (1995, p. 407) take it for granted that the burden is on the sympathizer of the description theory of reference to make her case. For argument's sake, I shall go along with S&W and grant the 'causal-historical' approach to reference.

bodies. Plausibly, the same can be said with respect to the theoretical terms deployed by folk psychology. S&W's conclusion is that even though the connectionist networks that RS&G considered showed that folk psychology is wrong, that would lend no support whatsoever to the more radical eliminativist claim that folk psychological posits do not exist.

In my opinion we need not worry about which approach to the theory of reference (descriptive, or causal-historical) is correct. As we saw in chapter 6, Stich (1990) interprets the thesis of eliminativism as the claim that the theoretical terms of folk psychology fail to refer. I am happy to concede that by granting that interpretation, and in particular, a causal-historical theory of reference, a logical gap in RS&G's argument may remain to be filled. Stich (1996), nonetheless, changes his mind, and claims against S&W that the theory of reference is not the place to go to when trying to settle ontological disputes. The following quote reveals the reasons that drive Stich to disagree with his previous line of reasoning:

> In some situations, it is easier to get a grant or a promotion or to enhance one's reputation in the scientific community by announcing the discovery of a new entity or denying the existence of one previously claimed to exist. In other situations, it is more politically expedient to conclude that entities of a certain sort don't have some of the properties previously attributed to them and that experimental results or other phenomena can best be explained by attributing some rather different properties to those entities. Which conclusion the scientific community ultimately accepts may well be determined, in some cases, by factors like these. (Stich, 1996, p. 68)

Stich's latest twist in the eliminativist plot makes of him a 'social constructivist', or, as he prefers, a Quinean pragmatist (see Stich, 1996, pp. 52-9; p. 72). Stich's approach has been winning support, surprisingly, among notorious friends of eliminativism. So, Patricia Churchland claims that in order to determine whether the entities of a non-basic theory can be identified or not with the entities of a new scientific successor, the decision

> is influenced by a variety of pragmatic and social considerations. The whim of the central investigators, the degree to which confusion will result from retention of the old terms, the desire to preserve or to break with past habits of thought, the related opportunities for publicizing the theory, cadging for grants, and attracting disciples all enter into decisions concerning whether to claim identities and therewith retention or whether to make the more radical claim of displacement. (Churchland, 1986, pp, 283-84)

Although I don't feel sympathetic with Stich's 'constructivist' approach, and Churchland's 'pragmatic' considerations (see below), we may agree with them for argument's sake that semantics cannot settle ontological disputes. Nevertheless, we need not worry about S&W's considerations. The purpose of this chapter is more modest in scope. I have tried to show that the thesis of propositional modularity is indeed inconsistent with a fully-superposed connectionist model of cognition. On the othe hand, I am happy to acknowledge the existence of S&W's aforementioned logical gap. Something else beyond the above incompatibility must be put forward in order to bring about the eliminativist conclusion. The results of this chapter, therefore, only represent a partial victory for the eliminativist. I believe however

that S&W's logical gap can be filled by digging elsewhere, delivering thus the goods for the eliminativist. As I see the issue, we must ignore S&W's considerations concerning the theory of reference, as well as Stich's, and Churchland's pragmatic considerations. In fact, I believe that we may reply to S&W's anti-eliminativist argument by exploiting a formal criterion to bridge the logical gap. A criterion that will help us determine *objectively* whether the posits of a discredited theory deserve to be eliminated, rather than retained. In Calvo Garzón (in preparation c) I argue that a reformulation of the reductionist/anti-reductionist debate, and in particular, a new challenging view of *intertheoretic reduction* developed by John Bickle in his recent *Psychoneural Reduction: The New Wave*, can help the eliminativist to complete her argument. But I must leave those matters for another occasion.

This chapter ends Part II of my dissertation. The careful reader will have noticed the existence of a common thread underlying both my connectionist defence of Quine's inscrutability thesis (Part I, chapters 4, and 5), and my connectionist defence of the elimination of the mental (Part II, chapter 7). Succintly, if the 'mind' of a cognitive agent is a fully-superposed connectionist network, then we shall not find discrete analogues of the words employed in a propositional ascription in the connectionist processing of the agent's cognitive system. Thus, the word 'gavagai ', or the belief *that p* will not have discrete connectionist analogues since, as we saw, those items are represented by means of highly idiosyncratic patterns of activation that defy a context-free symbolic treatment. According to the eliminative materialist mental states ought to be eliminated from our ontology. Among others, notorious

philosophers that have endorsed this position are Feyerabend (1963), Rorty (1970), and Churchland (1981). The natural sciences, according to Feyerabend, Rorty, or Churchland, have provided manifestable evidence in support of the view that *propositional* content does not exist. In this respect, I must emphasize that the results of my dissertation only serve to back a moderate form of eliminativism. Ultimately, the target, in my opinion, should be the elimination of *content*, rather than the elimination of propositional content. Most connectionist theorists nowadays prefer to frame the classical/connectionist debate as a debate about the architecture of cognition. Both sides assume a representationalist framework. Content, in the connectionist guise, is non-conceptual (i.e., does not conform to the conceptual patterns of classical constituency, and processing; see chapter 4 above).

For strategical reasons, I've assumed for the purposes of the current work a representationalist framework in connectionist theory. The current debate is an exciting one, and a lot is yet to be said. Nevertheless, I believe that the real significance of connectionist theory for the philosophy of language, and the philosophy of mind has not been fully appraised yet. Quine's thesis of the inscrutability of reference, for instance, aims at the right target. However, the way in which Quine tries to defend the thesis is not the most fortunate, and runs the risk of missing its real significance. Quine's strategy is to find more than one scheme of reference that fits all possible evidence (see chapter 1). However, the friend of semantic scepticism has a faster route to accomplish her task. It is not the fact that there is more than one correct theory of reference what threatens semantics. Rather, it is the fact that there is no semantic relation of reference *at all* between a speakers'

cognitive processes, and the external world. In chapter 4 we saw how to interpret semantically the space defined by the hidden units of a simple feedforward network. I think, nonetheless, that that furnishes us with a naive interpretation of connectionist networks. Put bluntly, I believe that every single pattern of behaviour (non-cognitive, as well as cognitive) is to be seen as mere causal correlations between inner states, and certain environmental features. Causal correlations that obviously do not suffice to establish *representational* status (cf., for example, Haugeland, 1991). To illustrate, we may say that humans are equivalent to sunflowers. The latter chase the sun, but noone would claim that they posses an inner representation of the sun. A full physical explanation in terms of causes and effects in real time, and real space, suffices to explain the sun-chasing behaviour of the sunflower. I believe that the same goes when we try to explain higher cognitive abilities. The patterns of behaviour to be explained are more complex *quantitatively*, but the principles are the same: causal correlations in the physical world. I just fail to see where the notion of representation can fit in this picture. I am aware that this is a radical claim, but unfortunately I still lack the conceptual, and technical apparatus to flesh out these thoughts.[35] That is a project that I hope I can take up soon, and produce a connectionist defence of a *General Theory of Anti-Representationalism*.

---

[35] The reader may care to consult Beer (1995a); Brooks (1991); Keijzer (1998); Port and Van Gelder (1995); Ramsey (1997); Thelen and Smith (1994); and van Gelder (1995) for some pioneer research in robotics, and dynamical systems that focuses upon a theory of anti-representationalism.

# References


Aizawa, K. (1997) 'Explaining Systematicity', *Mind and Language* 12, pp. 115-36.

Alston, W.P. (1996) *A Realist Conception of Truth*, Ithaca and London: Cornell University Press.

Arbib, M.A. (ed.) (1995) *The Handbook of Brain Theory and Neural Networks*, Cambridge, Mass.: MIT Press.

Armstrong, D.M. (1968) *A Materialist Theory of the Mind*, New York: Humanities Press.

Ayer, A.J. (1936) *Language, Truth and Logic*, London: Victor Gollancz.

Barrett, R. and Gibson, R. (1989) *Perspectives on Quine*, Oxford: Blackwell.

Barsalou, L.W. (1989) 'Intra-concept Similarity and Its Implications for Inter-concept Similarity' in Vosniadou, S. and Ortony, A. (eds.) (1989) *Similarity and Analogy*, Cambridge: Cambridge University Press.

Bechtel, P. (1980) 'Indeterminacy and Underdetermination: Are Quine's Two Theses Consistent?', *Philosophical Studies* 38, pp. 309-20.

Bechtel, W. (1991) 'Connectionism and the Philosophy of Mind: An Overview' in

Horgan, T. and Tienson, J. (eds.) (1991) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic Plublishers.

Bechtel, W. and Abrahamsen, A. (1991) *Connectionism and the Mind: And Introduction to Parallel Processing in Networks*, Oxford: Blackwell.

Beer, R.D. (1995a) 'Computational and Dynamical Languages for Autonomous Agents', in Port, R. and van Gelder, T. (1995) *Mind as Motion*, Cambridge, Mass.: MIT Press.

Beer, R.D. (1995b) 'A Dynamical Systems Perspective on Agent-Environment Interaction', *Artificial Intelligence* 72, pp. 173-215.

Benacerraf, P. (1983) 'What Numbers Could Not Be' in Benacerraf, P. and Putnam, H. (eds.) (1983) *Philosophy of Mathematics: Selected Readings*, Cambridge: Cambridge University Press.

Bermúdez, J.L. (1995a) 'Syntax, Semantics and Levels of Explanation', *Philosophical Quarterly* 45, pp. 361-67.

Bermúdez, J.L. (1995b) 'Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States', *Mind and Language* 10, pp. 333-69.

Bickle, J. (1989) *Towards a Contemporary Reformulation of the Mind-Body Problem*, Ph. D. Dissertation, University of California, Irvine.

Bickle, J. (1992) 'Revisionary Physicalism', *Biology and Philosophy* 7, pp. 411-30.

Bickle, J. (1993) 'Connectionism, Eliminativism, and The Semantic View of Theories', *Erkenntnis* 39, pp. 359-82.

Bickle, J. (1998) *Psychoneural Reduction: The New Wave*, Cambridge, Mass.: MIT Press.

Blackburn, S. (1984) *Spreading the Word*, Oxford: Oxford University Press.

Boden, M.A. (1990) *The Philosophy of Artificial Intelligence*, Oxford: Oxford

University Press.

Bogdan, R.J. (1993) 'The Architectural Nonchalance of Commonsense Psychology', *Mind and Language* 8, pp. 189-205.

Boghossian, P. (1990) 'The Status of Content', *Philosophical Review* 99, pp. 157-84.

Borges, J.L. (1971) *Ficciones*, Madrid: Alianza Editorial (First published in 1944 by Emecé Editores).

Bradley, M.C. (1975) 'Kirk on Indeterminacy of Translation', *Analysis* 36, pp. 18-22.

Bradley, M.C. (1976) 'Quine's Arguments for the Indeterminacy Thesis', *Australasian Journal of Philosophy* 54, pp. 24-49.

Brent, M.R. (1996) 'Advances in the Computational Study of Language Acquisition', *Cognition* 61, pp. 1-38.

Brooks, R. A. (1991) 'Intelligence Without Representation', *Artificial Intelligence* 47, pp. 139-59.

Calvo Garzón, F. (2000a) 'A Connectionist Defence of the Inscrutability Thesis', *Mind and Language*.

Calvo Garzón, F. (2000b) 'State Space Semantics and Conceptual Similarity: Reply to Churchland', *Philosophical Psychology* 13, pp. 77-95.

Calvo Garzón, F. (under review a) 'Semantic Perversity', *Teorema*.

Calvo Garzón, F. (under review b) 'Is Simplicity Alethic for Semantic Theories?', *Analysis*.

Calvo Garzón, F. (in preparation a) 'Can We Turn a Blind Eye to Eliminativism?'.

Calvo Garzón, F. (in preparation b) 'The Twilight of Propositional Content'.

Calvo Garzón, F. (in preparation c) 'Revising *Revisionary* Physicalism'.

Campbell, J. (1982) 'Knowledge and Understanding', *Philosophical Quarterly* 32, pp. 17-34.

Campbell, K. (1993) 'What Motivates Eliminativism', *Mind and Language* 8, pp. 206-10.

Cartwright, N. (forthcoming) *The Dappled World: Essays on the Perimeters of Science.*

Chappell, V.C. (ed.) (1962) *The Philosophy of Mind*, Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*, Cambridge, Mass.: MIT Press.

Chomsky, N. (1968) *Language and Mind*, New York: Harcourt, Brace and World.

Chomsky, N. (1969) 'Quine's Empirical Assumptions' in Davidson, D. and Hintikka, J. (eds.) (1969) *Words and Objections*, Dordrecht: Reidel.

Chomsky, N. (1980) *Rules and Representations*, Oxford: Blackwell.

Christiansen, M.H. and Chater, N. (1992) 'Connectionism, Learning and Meaning, *Connection Science* 4, pp. 227-52.

Churchland, P.M. (1981) 'Eliminative Materialism and the Propositional Attitudes', *Journal of Philosophy* 78; 2, pp. 67-90.

Churchland, P.M. (1984) *Matter and Consciousness*, Cambridge, Mass.: MIT Press.

Churchland, P.M. (1986) 'Some Reductive Strategies in Cognitive Neurobiology', *Mind* 95, pp. 279-309.

Churchland, P.M. (1989a) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge, Mass.: MIT Press.

Churchland, P.M. (1989b) 'On the Nature of Theories: A Neurocomputational Perspective' in Savage, C.W. (ed.) (1989) *The Nature of Theories:*

*Minnesota Studies in the Philosophy of Science* XIV, Minneapolis: University of Minnesota Press.

Churchland, P.M. (1990) 'On the Nature of Explanation: A PDP Approach', *Physica D* 42, pp. 281-92.

Churchland, P.M. (1993) 'Evaluating Our Self Conception', *Mind and Language* 8, pp. 211-22.

Churchland, P.M. (1995) *The Engine of Reason, The Seat of the Soul*, Cambridge, Mass.: MIT Press.

Churchland, P.M. (1996) 'Learning and Conceptual Change: The View from the Neurons' in Clark, A. and Millican, P.J.R. (1996) *Connectionism, Concepts and Folk Psychology: The Legacy of Alan Turing*, Vol. 2, Oxford: Clarendon Press.

Churchland, P.M. (1998) 'Conceptual Similarity Across Sensory and Neural Diversity: the Fodor/Lepore Challenge Answered', *The Journal of Philosophy* 95, pp. 5-32.

Churchland, P.S. (1986) *Neurophilosophy: Toward a Unified Science of the Mind-Brain*, Cambridge, Mass.: MIT Press./Bradford Books.

Churchland, P.S. and Sejnowski, T. (1992) *The Computational Brain*, Cambridge, Mass.: MIT Press

Clark, A. (1989) *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, Mass.:MIT Press.

Clark, A. (1991a) 'Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn' in Horgan, T. and Tienson, J. (eds.) (1991) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic Plublishers.

Clark, A. (1991b) 'Radical Ascent', *The Aristotelian Society* 65 (supplement).

Clark, A. (1993a) *Associative Engines*, Cambridge, Mass.: MIT Press.

Clark, A. (1993b) 'The Varieties of Eliminativism: Sentential, Intentional and Catastrophic', *Mind and Language* 8, pp. 223-33.

Clark, A. (1994a) 'Representational Trajectories in Connectionist Learning', *Minds and Machines* 4, pp. 317-32.

Clark, A. (1994b) 'Language of Thought (2)' in Guttenplan, S. (ed.) (1994) *A Companion to the Philosophy of Mind*, Oxford: Blackwell.

Clark, A. (1995) 'Connectionist Minds' in Macdonald and, C. and Macdonald, G., (eds.) (1995b) *Connectionism: Debates on Psychological Explanation*, Vol. 2, Oxford: Blackwell.

Clark, A. (1997) *Being There*, Cambridge, Mass.: MIT Press.

Clark, A. and Chalmers, D. (unpublished) 'The Extended Mind'.

Clark, A. and Lutz, R. (eds.) (1992) *Connectionism in Context*, London: Springer-Verlag.

Clark, A. and Millican, P.J.R. (1996) *Connectionism, Concepts and Folk Psychology: The Legacy of Alan Turing*, Vol. 2, Oxford: Clarendon Press.

Clark, A. and Toribio, J. (1994) 'Doing Without Representing?', *Synthese*, pp. 401-31.

Cleeremans, A. (1993) *Mechanisms of Implicit Learning*, Cambridge, Mass.: MIT Press.

Cohen, R.S. et al. (eds.) (1976) *Essays in Memory of Imre Lakatos*, Dordrecht-Holland: Reidel Publishing Company.

Collins, A. and Quillian, M. (1972) 'Experiments on Semantic Memory and Language Comprehension' in Gregg, L. (ed.) (1972) *Cognition in Learning*

*and Memory*, New York: Wiley.

Cottrell, G.W. (1987) 'Toward Connectionist Semantics', *Theoretical Issues in Natural Language Processing* 3, pp. 65-70. New Mexico: Association for Computational Linguistics, University of New Mexico.

Cummins, R. (1989) *Meaning and Mental Representation*, Cambridge, Mass.: MIT Press.

Cummins, R. (1991) 'Form, Interpretation, and the Uniqueness of Content: Response to Morris', *Minds and Machines* 1, pp. 31-42.

Cummins, R. (1996) 'Systematicity', *Journal of Philosophy* 93, pp. 591-614.

Cussins, A. (1990) 'The Connectionist Construction of Concepts' in Boden, M.A. (1990) *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.

Cussins, A. (1993) 'Nonconceptual Content and the Elimination of Misconceived Composites!', *Mind and Language* 8, pp. 234-52.

Dancy, J. and Sosa, E. (1992) *A Companion to Epistemology*, Oxford: Blackwell.

Davidson, D. (1980a) *Essays on Actions and Events*, Oxford: Oxford University Press.

Davidson, D. (1980b) 'Mental Events' in Davidson (1980a), pp. 207-25.

Davidson, D. (1984) *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press.

Davidson, D. and Harman, G. (eds.) (1972) *Semantics of Natural Language*, Dordrecht: Reidel.

Davidson, D. and Hintikka, J. (eds.) (1969) *Words and Objections*, Dordrecht: Reidel.

Davies, M. (1986) 'Tacit Knowledge and the Structure of Thought and Language'

in Travis, C. (ed.) (1986) *Meaning and Interpretation*, Oxford: Blackwell.

Davies, M. (1987) 'Tacit Knowledge and Semantic Theory: Can a Five per cent Difference Matter?', *Mind* 96, pp. 441-62.

Davies, M. (1989a) 'Tacit Knowledge and Subdoxatic States' in George, A. (1989) *Reflections on Chomsky*, Oxford: Blackwell.

Davies, M. (1989b) 'Connectionism, Modularity and Tacit Knowledge', *British Journal for the Philosophy of Science* 40, pp. 541-55.

Davies, M. (1991) 'Concepts, Connectionism and the Language of Thought' in Ramsey, W., Stich, S. and Rumelhart, D.E. (eds.) (1991) *Philosophy and Connectionist Theory*, Hillsdale, N.J.: Lawrence Erlbaum.

Davies, M. (1995) 'Two Notions of Implicit Rules' in Tomberlin, J. (ed.) (1995) *Philosophical Perspectives*, 9: AI, Connectionism and Philosophical Psychology, Ridgeview.

Davis, S. (1992) *Connectionism: Theory and Practice*, New York: Oxford University Press.

Divers, J. and Miller, A. (1994) 'Why Expresivists about Value Should Love Minimalism about Truth', *Analysis* 54, pp. 12-9.

Divers, J. and Miller, A. (1995) 'Platitudes and Attitudes: A Minimalist Conception of Belief', *Analysis* 55, pp. 37-44.

Dorffner, (1992) 'A Step Toward Sub-Symbolic Language Models Without Linguistic Representations' in Reilly, R. and Sharkey, N.E. (eds.) (1992) *Connectionist Approaches to Natural Language Processing*, Hove: Erlbaum.

Dretske, F (1988) *Explaining Behavior: Reasons in a World of Causes*, Cambridge, Mass.: MIT Press.

Dummett, M. (1974a) 'The significance of Quine's Indeterminacy Thesis', *Synthese* 27, pp. 351-97.

Dummett, M. (1974b) 'Reply to W.V. Quine', *Synthese* 27, pp. 413-16.

Dummett, M. (1975) 'What is a Theory of Meaning' in Guttenplan, S. (ed.) (1975) *Mind and Language*, Oxford: Oxford University Press.

Dummett, M. (1976) 'What is a Theory of Meaning? (II)' in Evans, G. and McDowell, J. (eds.) (1976) *Truth and Meaning*, Oxford: Clarendon Press.

Elman, J. (1989) 'Structured Representations and Connectionist Networks', *CRL Technical Report* 8901, San Diego, CA: University of California.

Elman, J. (1990) 'Finding Structure in Time', *Cognitive Science* 14, pp. 179-211.

Elman, J. (1992) 'Grammatical Structure and Distributed Representations' in Davis, S. (1992) *Connectionism: Theory and Practice*, New York: Oxford University Press.

Elman, J. (1998) 'Generalization, Simple Recurrent Networks, and the Emergence of Structure' in M.A. Gernsbacher and S. Derry (eds.) *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, Mahway, NJ: Lawrence Erlbaum.

Elman, J. et al. (1996) *Rethinking Innateness*, Cambridge, Mass.: MIT Press.

Edwards, J. (1994) 'Critical Notice: Debates About Realism Transposed to a New Key', *Mind* 103, pp. 59-72.

Evans, G. (1975) 'Identity and Predication', *Journal of Philosophy* 72, pp. 343-63.

Evans, G. (1981) 'Semantic Theory and Tacit Knowledge' in Holtzmann, S. and Leich, C. (eds.) (1981) *Wittgenstein: To Follow a Rule*, London: Routledge.

Evans, G. (1982) *The Varieties of Reference*, Oxford: Oxford University Press.

Evans, G. (1985) *Collected Papers*, Oxford: Oxford University Press.

Evans, G. and McDowell, J. (eds.) (1976) *Truth and Meaning*, Oxford: Clarendon Press.

Evnine, S. (1991) *Donald Davidson*, Oxford: Polity Press (in association with Blackwell).

Eysenck, M.W. and Keane, M.T. (1995) *Cognitive Psychology: A Student's Handbook*, Erlbaum: Psychology Press.

Feyerabend, P. (1963) 'Materialism and the Mind-Body Problem', *Review of Metaphysics* 17, pp. 49-66.

Field, H. (1972) 'Tarski's Theory of Truth', *Journal of Philosophy* 69, pp. 347-75.

Field, H. (1973) 'Theory Change and the Indeterminacy of Reference', *Journal of Philosophy* 70, pp. 462-81.

Field, H. (1974) 'Quine and the Correspondence Theory', *Philosophical Review* 83, pp. 200-28.

Field, H. (1975) 'Conventionalism and Instrumentalism in Semantics', *Nous* 9, pp. 375-405.

Field, H. (1978) 'Mental Representations', *Erkenntnis* 13, pp. 9-61.

Field, H. (1980) *Science without Numbers*, Oxford: Basil Blackwell.

Fodor, J. (1975) *The Language of Thought*, Hassocks: Harvester Press.

Fodor, J. (1981) *Representations*, Cambridge, Mass.: MIT Press.

Fodor, J. (1987) *Psychosemantics*, Cambridge, Mass.: MIT Press.

Fodor, J. (1990a) *A Theory of Content and Other Essays*, Cambridge, Mass.: MIT Press.

Fodor, J. (1990b) 'Making Mind Matter More' in Fodor (1990a).

Fodor, J. (1994) *The Elm and the Expert: Mentalese and its Semantics*, Cambridge, Mass.: MIT Press.

Fodor, J. (1998a) *Concepts. Where Cognitive Science Went Wrong*, Oxford: Clarendon Press.

Fodor, J. (1998b) *In Critical Condition*, Cambridge, Mass.: MIT Press.

Fodor, J. and Lepore, E. (1992) *Holism: A Shopper's Guide*, Oxford: Blackwell.

Fodor, J. and Lepore, E. (1996) 'Paul Churchland and State Space Semantics' in McCauley, R. (ed.) (1996) *The Churchlands and Their Critics*, Oxford: Blackwell.

Fodor, J. and Lepore, E. (forthcoming) 'All At Sea in Semantic Space: Churchland on Meaning Similarity', *The Journal of Philosophy.*

Fodor, J. and McLaughlin, B. (1990) 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work', *Cognition* 35, pp. 183-204.

Fodor, J. and Pylyshyn, Z. (1988) 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition* 28, pp. 3-71.

Føllesdal, D. (1975) 'Meaning and Experience' in Guttenplan, S. (ed.) (1975) *Mind and Language*, Oxford: Oxford University Press.

Føllesdal, D. (1989) 'Indeterminacy and Mental States' in Barrett, R. and Gibson, R. (1989) *Perspectives on Quine*, Oxford: Blackwell.

Forster, M. and Saidel, E. (1994) 'Connectionism and the Fate of Folk Psychology: A Reply to Ramsey, Stich and Garon', *Philosophical Psychology* 7, pp. 437-52.

Foster, L. and Swanson J.W. (eds.) (1970) *Experience and Theory*, London: Duckworth.

Freeman, W.J. and Skarda, C.A. (1990) 'Representations: Who Needs Them?', in McGaugh, J.L., Weinberger, J.L., and Lynch, G. (eds.) (1990) *Brain*

*Organization and Memory Cells, Systems and Circuits*, Guildford Press.

Friedman, M. (1975) 'Physicalism and the Indeterminacy of Translation', *Nous* 9, pp. 353-74.

García-Carpintero, M. (1995) 'The Philosophical Import of Connectionism: A Critical Notice of Andy Clark's *Associative Engines*', *Mind and Language* 10, pp. 370-401.

García Márquez, G. (1997) *Cien Años de Soledad*, Madrid: Ediciones Cátedra.

Gardner, M. (1973) 'Apparent Conflicts Between Quine's Indeterminacy Thesis and his Philosophy of Science', *British Journal for the Philosophy of Science* 24, pp. 381-93.

Garson, J.W. (1998) 'Chaotic Emergence and The Language of Thought', *Philosophical Psychology*, 11, pp. 303-15.

George, A. (1989) *Reflections on Chomsky*, Oxford: Blackwell.

Gibson, R. (1982) *The Philosophy of W.V. Quine*, Tampa: University Presses of Florida.

Godfrey-Smith, P. (1986) 'Why Semantic Properties Won't Eran Their Keep', *Philosophical Studies* 50, pp. 223-36.

Goodman, N. (1965) *Fact, Fiction and Forecast*, Indianapolis: Bobbs-Merrill.

Goodman, N. (1978) *Ways of Worldmaking*, Indianapolis: Hackett.

Gorman, R. and Sejnowski, T. (1988) 'Learned Classification of Sonar Targets ussing a Massively Parallel Network', *IEEE Transactions: Acoustics, Speech and Signal Processing* 36(7), pp. 1135-40.

Guttenplan, S. (ed.) (1975) *Mind and Language*, Oxford: Oxford University Press.

Guttenplan, S. (ed.) (1994) *A Companion to the Philosophy of Mind*, Oxford: Blackwell.

Haack, S. (1978) *Philosophy of Logics*, Cambridge: Cambridge University Press.

Haas, W. (1968) 'The Theory of Translation' in Parkinson, G.H.R. (1968) *The Theory of Meaning*, Oxford: Oxford University Press.

Hacking, I. (1975) *Why Does Language Matter to Philosophy?*, Cambridge: Cambridge University Press.

Hadley, R.F. (1992) 'Compositionality and Systematicity in Connectionist Language Learning', *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 659-670.

Hadley, R.F. (1997) 'Cognition, Systematicity and Nomic Necessity', *Mind and Language* 12, pp. 137-53.

Hale, B. and Wright, C. (eds.) (1997) *A Companion to the Philosophy of Language*, Oxford: Blackwell.

Hannan, B. (1993) 'Don't Stop Believing: The Case Against Eliminative Materialism', *Mind and Language* 8, pp. 165-79.

Hanson, N.R. (1958) *Patterns of Discovery*, Cambridge: Cambridge University Press.

Hanson, S.J. and Burr, D.J. (1990) 'What Connectionist Models Learn: Learning and Representation in Connectionist Networks', *Behavioral and Brain Sciences*, 13, pp. 471-518.

Hare, R.M. (1952) *The Language of Morals*, Oxford: Oxford University Press.

Harman, G. (1969) 'An Introduction to "Translation and Meaning" Chapter Two of Word and Object' in Davidson, D. and Hintikka, J. (eds.) (1969) *Words and Objections*, Dordrecht: Reidel.

Harman, G. (1973) *Thought*, Princeton, NJ: Princeton University Press.

Harnad, S. (1990) 'The Symbol Grounding Problem', *Physica D* 42, pp. 335-46.

Haugeland, J (1981) *Mind Design*, Cambridge, Mass.: MIT Press.

Haugeland, J (1991) 'Representational Genera' in Ramsey, W., Stich, S. and Rumelhart, D.E. (eds.) (1991) *Philosophy and Connectionist Theory*, Hillsdale, N.J.: Lawrence Erlbaum.

Hecht-Nielsen, R. (1989) 'Theory of the Backpropagation Neural Network', in *Proceedings of the ational Joint Conference on Neural Networks* (Washington, DC), vol. I, 593-605. Piscataway, NJ: IEEE.

Heil, J. and Mele, A. (1993) *Mental Causation*, Oxford: Oxford Univ. Press.

Hendler, J. (1990) 'But What is the Substance of Connectionist Representation?', *Behavioral and Brain Sciences*, 13, pp. 496-97.

Hendriks-Jansen, H. (1996) *Catching Ourselves in the Act*, Cambridge, Mass.:MIT Press.

Higginbotham, J. (1989) 'Knowledge of Reference' in George, A. (1989) *Reflections on Chomsky*, Oxford: Blackwell.

Hill, C. (1971) 'Gavagai', *Analysis* 32, pp. 68-75.

Hintikka, J. (1973) *Logic, Language-Games and Information*, Oxford: Clarendon Press.

Hintikka, J. (1976) 'Quantifiers in Logic and Quantifiers in Natural Languages' in Körner, S. (ed.) (1976) *Philosophy of Logic*, Oxford: Blackwell.

Hinton, G.E. and Anderson, J.A. (1981) *Parallel Models of Associative Memory*, Hillsdale, NJ: Erlbaum.

Holtzmann, S. and Leich, C. (eds.) (1981) *Wittgenstein: To Follow a Rule*, London: Routledge.

Hooker, C.A. (1981) 'Towards a General Theory of Reduction. Part I: Historical

and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction', *Dialogue* 20, pp. 38-59; 201-36; 496-529.

Hooker, C.A. (1995) *Reason, Regulation, and Realism: Towards a Regulatory Systems Theory of Reason and Evolutionary Epistemology*, Albany: State University of New York Press.

Hookway, C. (1978) 'Indeterminacy and Interpretation' in Hookway, C. and Pettit, P. (eds.) (1978) *Action and Interpretation*, Cambridge: Cambridge University Press.

Hookway, C. (1988) *Quine*, Oxford: Polity Press.

Hopcroft, J.E., & Ullman, J.D. (1979). Introduction to automata theory, languages, and computation. Reading, MA: Addison-Wesley. Pp. 42-44.

Horgan, T. and Graham, G. (1991) 'In Defense of Southern Fundamentalism', *Philosophical Studies* 62, pp. 107-34.

Horgan, T. and Tienson, J. (eds.) (1991) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic Plublishers.

Horgan, T. and Tienson, J. (1992) 'Structured Representations in Connectionist Systems?' in Davis, S. (1992) *Connectionism: Theory and Practice*, New York: Oxford University Press.

Horgan, T. and Tienson, J. (1994) 'Representations Don't Need Rules: Reply to James Garson', *Mind and Language* 9, pp. 38-55.

Horgan, T. and Tienson, J. (1996) *Connectionism and the Philosophy of Psychology*, Cambridge, Mass.: MIT Press.

Hornik, K., Stinchcombe, M., and White, H. (1989) 'Mulltilayer Feedforward Networks are Universal Approximators', *Neural Networks* vol. 2, 359-66.

Horwich, P. (1990) *Truth*, Oxford: Basil Blackwell.

Jackson, F. (1994) 'Realism, Truth and Truth Aptness', *Philosophical Books* 35, pp. 162-9.

Jackson, F., Oppy, G. and Smith, M. (1994) 'Minimalism and Truth Aptness', *Mind* 103, pp. 287-302.

Jackson, F. and Pettit, P. (1988) 'Functionalism and Broad Content', *Mind* 97, pp. 381-400.

Jackson, F. and Pettit, P. (1990) 'In Defense of Folk Psychology', *Philosophical Studies* 59, pp. 31-54.

Jackson, F. and Pettit, P. (1993) 'Folk Belief and Commonplace Belief', *Mind and Language* 8, pp. 298- 305.

Jordan, M.I. (1986) 'Attractor Dynamics and Parallelism in a Connectionist Sequential Machine' in *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, pp. 10-7.

Keijzer, F.A. (1998) 'Doing Without Representations Which Specify What To Do', *Philosophical Psychology*, 11, pp. 269-302.

Kim, J. (1996) *Philosophy of Mind*, Oxford: Westview Press.

Kirk, R. (1969) 'Translation and Indeterminacy', *Mind* 78, pp. 321-41.

Kirk, R. (1973) 'Underdetermination of Theory and Indeterminacy of Translation', *Analysis* 33, pp. 195-202.

Kirk, R. (1977) 'More on Quine's Reasons for Indeterminacy of Translation', *Analysis* 37, pp. 136-41.

Kirk, R. (1979) 'From Physical Explicability to Full-Blooded Materialism', *Philosophical Quarterly* 20, pp. 229-37.

Kirk, R. (1982) 'On Three Alledged Rivals to Homophonic Translation', *Philosophical Studies* 42, pp. 409-18.

Kirk, R. (1983) 'Quinean Indeterminacy and Forcing', *Erkenntnis* 20, pp. 213-18.

Kirk, R. (1985) 'Davidson and Indeterminacy of Translation', *Analysis* 45, pp. 20-4.

Kirk, R. (1986) *Translation Determined*, Oxford: Oxford University Press.

Kripke, S. (1972) 'Naming and Necessity' in Davidson, D. and Harman, G. (eds.) (1972) *Semantics of Natural Language*, Dordrecht: Reidel.

Kripke, S. (1982) *Wiitgenstein on Rules and Private Language*, Oxford: Basil Blackwell.

Laakso, A. and Cottrell, G.W. (1998) 'How Can *I* Know What *You* Think?: Assessing Representational Similarity in Neural Systems', *Proceedings of the 20th Annual Cognitive Science Conference*, Madison, WI. Mahway: Lawrence Erlbaum.

Laakso, A. and Cottrell, G.W. (2000) 'Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems', *Philosophical Psychology.*

Leonardi, P. and Santambrogio, M. (eds.) (1995) *On Quine: New Essays*, Cambridge: Cambridge University Press.

Levy, E. (1970) 'Competing Radical Translation: Examples, Limitations and Implications', *Boston Studies in the Philosophy of Science* 8, pp. 590-605.

Lewis, D. (1972) 'Psychophysical and Theoretical Identification', *Australasian Journal of Philosophy*, 50; 3, pp. 247-58.

Lewis, D. (1974) 'Radical Interpretation', *Synthese* 27, pp. 331-44.

Loewer, B. (1997) 'A Guide to Naturalizing Semantics' in Hale, B. and Wright, C. (eds.) (1997) *A Companion to the Philosophy of Language*, Oxford: Blackwell.

Loewer, B. and Rey, G. (1991) *Meaning in Mind: Fodor and His Critics*, Oxford:

Blackwell.

Macdonald, C. (1989) *Mind-Body Identity Theories*, London: Routledge.

Macdonald, C. (1997) 'Connectionism and Eliminativism: Reply to Stephen Mills', *International Journal of Philosophical Studies* 5, pp. 316-22.

Macdonald, C. and Macdonald, G. (eds.) (1995a) *Philosophy of Psychology: Debates on Psychological Explanation*, Vol. 1, Oxford: Blackwell.

Macdonald, C. and Macdonald, G. (eds.) (1995b) *Connectionism: Debates on Psychological Explanation*, Vol. 2, Oxford: Blackwell.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*, Harmondsworth: Penguin Books.

Marcus, G. (1998) 'Symposium on Cognitive Architecture: The Algebraic Mind', in M.A. Gernsbacher & S. Derry (eds.) *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates, p. 6.

Marr, D. (1982) *Vision*, New York: W.H. Freeman.

Massey, G. (1978) 'Indeterminacy, Inscrutability and Ontological Relativity', *Studies in Ontology, American Philosophical Quaterly*, Monograph Series No 12, pp. 43-55.

McCauley, R. (ed.) (1996) *The Churchlands and Their Critics*, Oxford: Blackwell.

McClelland, J.L., Rumelhart, D.E. and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, Cambridge, Mass,: MIT Press/Bradford Books.

McClelland, J.L. and Kawamoto, A.H. (1986) 'Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences' in McClelland, Rumelhart et al., pp. 272-325.

McCulloch, G. (1995) *The Mind and its World*, London and New York: Routledge.

McDowell, J. (1976) 'Truth Conditions, Bivalence and Verificationism' in Evans, G. and McDowell, J. (eds.) (1976) *Truth and Meaning*, Oxford: Clarendon Press.

McDowell, J. (1981) 'Anti-realism and the Epistemology of Understanding' in Parret, H. and Bouveresse, J. (1981) *Meaning and Understanding*, Berlin; New York: Walter de Gruyter.

McGinn, C. (1982) *The Character of Mind*, Oxford: Oxford Univ. Press.

Miikkulainen, R. (1993) *Subsymbolic Natural Language Processing. An Integrated Model of Scripts, Lexicon, and Memory*, Cambridge, MA: MIT Press.

Miller, A. (1997a) 'Tacit Knowledge' in Hale, B. and Wright, C. (eds.) (1997) *A Companion to the Philosophy of Language*, Oxford: Blackwell.

Miller, A. (1997b) *Philosophy of Language*, London: University College London Press.

Millikan, R. (1984) *Language, Thought and Other Biological Categories*, Cambridge, Mass.: MIT Press.

Millikan, R. (1994) 'Biosemantics' in Stich, S and Warfield, T (1994) *Mental Representation: A Reader*, Oxford: Blackwell.

Mills, S. (1997) 'Connectionism: Debates on Psychological Explanation', *International Journal of Philosophical Studies* 5, pp. 95-110.

Minsky, M. (1985) *Society of Mind*, New York: Simon & Schuster.

Minsky, M. and Papert, S. (1988) *Perceptrons: An Introduction to Computational Geometry*, Cambridge, MA: MIT Press.

Morris, M. (1991) 'Why There Are No Mental Representations', *Minds and Machines* 1, pp. 1-30.

Moser, P.K. and Trout, J.D. (eds.) (1995) *Contemporary Materialism: A Reader*, London and New York: Routledge.

Nagel, E. (1961) *The Structure of Science*, New York: Harcourt, Brace, and World.

Nerlich, G. (1976) 'Quine's "Real Ground"', *Analysis* 37, pp. 15-9.

Newell, A. (1980) 'Physical Symbol Systems', *Cognitive Science*, 4, pp. 135-83.

Newell, A. and Simon, H. (1972) *Human Problem Solving*, Englewood Cliffs, NJ: Prentice Hall.

Niklasson, L.F. and Van Gelder, T. (1994) 'On Being Systematically Connectionist', *Mind and Language* 9, pp. 288-302.

Ondaatje, M. (1993) *The English Patient*, London: Picador.

Oppenheim, P. and Putnam, H. (1958) 'Unity of Science as a Working Hypothesis' in H. Feigl, M. Scriven, and G. Maxwell (eds.) *Concepts, Theories, and the Mind-Body Problem*, Minnesota Studies in the Philosophy of Science, 2, pp. 3-36. Minneapolis: University of Minnesota Press.

Parker, D.B. (1985) 'Learning Logic', *Technicl Report TR-47*, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA.

Parker, D.B. (1987) 'Optimal Algorithms for Adaptive Networks: Second Order Back Propagation, Second Order Direct Propagation, and Second Order Hebbian Learning', in *Proceedings of the IEEE First International Conference on Neural Networks* (San Diego, CA), vol. II, 593-600. Piscataway, NJ: IEEE.

Parkinson, G.H.R. (1968) *The Theory of Meaning*, Oxford: Oxford University Press.

Parret, H. and Bouveresse, J. (1981) *Meaning and Understanding*, Berlin; New

York: Walter de Gruyter.

Peacocke, C. (1989) 'When is a Grammar Psychologically Real?' in George, A. (1989) *Reflections on Chomsky*, Oxford: Blackwell.

Peacocke, C. (1992) *A Study of Concepts*, Cambridge, Mass.: MIT Press.

Place, U.T. (1992) 'Eliminative Connectionism –Its Implications for a Return to an Empiricist Behaviorist Linguistics', *Behavior and Philosophy* 20, pp. 21-35.

Platts, M.B. (1979) *Ways of Meaning: An Introduction to A Philosophy of Language*, London: Routledge and Kegan Paul.

Plunkett. K. and Elman, J. (1997) *Exercises in Rethinking Innateness*, Cambridge, Mass.: MIT Press.

Pollack, J. (1990) 'Recursive Distributed Representations', *Artificial Intelligence* 46, pp. 77-105.

Port, R. and van Gelder, T. (1995) *Mind as Motion*, Cambridge, Mass.: MIT Press.

Putnam, H. (1975) *Mind, Language and Reality*, Cambridge: Cambridge University Press.

Putnam, H. (1981a) *Reason, Truth and History*, Cambridge: Cambridge University Press.

Putnam, H. (1981b) 'Reductionism and the Nature of Mental States' in Haugeland, J. (1981) *Mind Design*, Cambridge, Mass.: MIT Press.

Putnam, H. (1983) *Realism and Reason*, Cambridge: Cambridge University Press.

Putnam, H. (1988) *Representation and Reality*, Cambridge, Mass.: MIT Press.

Quine, W.V. (1951) 'Semantics and Abstract Objects', *Proceedings of the American Academy of Arts and Sciences* 80, pp. 90-6.

Quine, W.V. (1953) *From a Logical Point of View*, Cambridge, Mass.: Harvard University Press.

Quine, W.V. (1960) *Word and Object*, Cambridge, Mass.: MIT Press.

Quine, W.V. (1963) *Set Theory and Its Logic*, Cambridge, Mass.: The Belknap Press.

Quine, W.V. (1969a) *Ontological Relativity and Other Essays*, New York: Columbia University Press.

Quine, W.V. (1969b) 'Reply to Chomsky' in Davidson, D. and Hintikka, J. (eds.) (1969) *Words and Objections*, Dordrecht: Reidel.

Quine, W.V. (1970a) *Philosophy of Logic*, Englewood Cliffs, N.J.: Prentice-Hall.

Quine, W.V. (1970b) 'Grades of Theoreticity' in Foster, L. and Swanson J.W. (eds.) (1970) *Experience and Theory*, London: Duckworth.

Quine, W.V. (1970c) 'Methodological Reflections in Current Linguistic Theory', *Synthese* 21, pp. 386-98.

Quine, W.V. (1970d) 'On the Reasons for Indeterminacy of Translation', *Journal of Philosophy* 67, pp. 178-83.

Quine, W.V. (1973) *The Roots of Reference*, La Salle, Ill.: Open Court.

Quine, W.V. (1975a) 'On Empirically Equivalent Systems of the World', *Erkenntnis* 9, pp. 313-28.

Quine, W.V. (1975b) 'The Nature of Natural Knowledge' in Guttenplan, S. (ed.) (1975) *Mind and Language*, Oxford: Oxford University Press.

Quine, W.V. (1975c) 'Mind and Verbal Dispositions' in Guttenplan, S. (ed.) (1975) *Mind and Language*, Oxford: Oxford University Press.

Quine, W.V. (1976) 'Whiter Physical Objects?' in Cohen, R.S. et al. (eds.) (1976) *Essays in Memory of Imre Lakatos*, Dordrecht-Holland: Reidel Publishing Company.

Quine, W.V. (1979a) 'Facts of the Matter' in Shahan, R.W. and Swoyer, C.V.

(eds.) (1979) *Essays on the Philosophy of W.V. Quine*, Hassocks: Harvester.

Quine, W.V. (1979b) 'Comment on Newton-Smith', *Analysis* 39, pp. 66-7.

Quine, W.V. (1981a) *Theories and Things*, Cambridge, Mass.: Harvard University Press.

Quine, W.V. (1981b) 'Five Milestones of Empiricism' in Quine (1981a).

Quine, W.V. (1987) 'Indeterminacy of Translation Again', *Journal of Philosophy* 84, pp. 5-10.

Quine, W.V. (1989) 'Three Indeterminacies' in Barrett, R. and Gibson, R. (1989) *Perspectives on Quine*, Oxford: Blackwell.

Quine, W.V. (1990) *Pursuit of Truth*, Cambridge, Mass.: Harvard University Press.

Ramsey, F. (1927) 'Facts and Propositions', *Proceedings of the Aristotelian Society*, (supplement).

Ramsey, W. (1992) 'Connectionism and the Philosophy of Mental Representation' in Davis, S. (1992) *Connectionism: Theory and Practice*, New York: Oxford University Press.

Ramsey, W. (1994) 'Distributed Representation and Causal Modularity: A Rejoinder to Forster and Saidel', *Philosophical Psychology* 7, pp. 453-61.

Ramsey, W. (1997) 'Do Connectionist Representations Earn Their Explanatory Keep?', *Mind and Language* 12, pp. 34-66.

Ramsey, W., Stich, S. and Garon, J. (1991) 'Connectionism, Eliminativism and the Future of Folk Psychology' in Ramsey, W., Stich, S. and Rumelhart, D.E. (eds.) (1991).

Ramsey, W., Stich, S. and Rumelhart, D.E. (eds.) (1991) *Philosophy and Connectionist Theory*, Hillsdale, N.J.: Lawrence Erlbaum.

Ray, G. (1997) 'Fodor and the Inscrutability Problem', *Mind and Language* 12, pp. 475-89.

Reber, A.S. (1967) 'Implicit Learning of Artificial Grammars', *Journal of Verbal Learning and Verbal Behavior*, 6, pp. 855-863.

Reber, A.S. (1976) 'Implicit Learning of Synthetic Languages: The Role of the Instructional Set', *Journal of Experimental Psychology: Human Learning and Memory*, 2, pp. 88-94.

Reilly, R. and Sharkey, N.E. (eds.) (1992) *Connectionist Approaches to Natural Language Processing*, Hove: Erlbaum.

Rey, G. (1991) 'An Explanatory Budget for Connectionism and Eliminativism' in Horgan, T. and Tienson, J. (eds.) (1991) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic Plublishers.

Romanos, G. (1983) *Quine and Analytical Philosophy*, Cambridge, Mass.: MIT Press.

Rorty, R. (1970) 'In Defense of Eliminative Materialism', *Review of Metaphysics* 24, pp. 112-21.

Rorty, R. (1972) 'Indeterminacy of Translation and of Truth', *Synthese* 23, pp. 443-62.

Rorty, R. (1979) *Philosophy and the Mirror of Nature*, Princeton: Princeton University Press.

Rosch, E.H. (1975) 'Cognitive Representations of Semantic Categories', *Journal of Experimental Psychology: General* 104, pp. 192-233.

Rosenberg, C. and Sejnowski, J. (1987) 'Parallel Networks that Learn to Pronounce English Text', *Complex Systems* 1, pp. 145-68.

Rosenblatt, F. (1959) *Principles of Neurodynamics*, New York: Spartan Books.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) 'Learning Internal Representations by Error Propagation' in Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Cambridge, Mass.: MIT Press/Bradford Books.

Rumelhart, D.E., McClelland, J.L. (1982) 'An Interactive Activation Model of Context Effects in Letter Perception: Part II. The Contextual Enhancement Effect and Some Tests and Extensions of the Model', *Psychological Review* 89, pp. 60-94.

Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Cambridge, Mass.: MIT Press/Bradford Books.

Schiffer, S. (1987) *Remnants of Meaning*, Cambridge, Mass.: MIT Press.

Schopman, J. and Shawky, A. (1996) 'Remarks on the Impact of Connectionism on Our Thinking about Concepts' in Clark, A. and Millican, P.J.R. (1996) *Connectionism, Concepts and Folk Psychology: The Legacy of Alan Turing*, Vol. 2, Oxford: Clarendon Press.

Schyns, P. (1991) 'A Modular Neural Network Model of Concept Acquisition', *Cognitive Science* 15, pp. 461-508.

Schyns, P., Goldstone, R. and Thibaut, J.P. (1998) 'The Development of Features in Object Concepts', *Behavioral and Brain Sciences*21, pp. 1-53.

Schyns, P. and Rodet, L. (1997) 'Categorization Creates Functional Features', *Journal of Experimental Psychology: Learning, Memory and Cognition* 23/3, pp. 1-16.

Sejnowski, T. and Rosenberg, C. (1986) 'NETtalk: A Parallel Network That Learns

to Read Aloud', Technical Report JHU/EECS-86/01, Johns Hopkins University.

Sellars, W. (1968) *Science and Metaphysics: Variations on Kantian Themes*, New York: Humanities Press.

Servan-Schreiber, D., Cleeremans, A. and Mc Clelland, J.L. (1988) 'Encoding Sequential Structure in Simple Recurrent Networks', Technical Report CMU- CS-88-183, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.

Servan-Schreiber, D., Cleeremans, A. and Mc Clelland, J.L. (1989) 'Learning Sequential Structure in Simple Recurrent Networks' in Toureztky, D.S. (ed.) (1989) *Advances in Neural Information Processing Systems I*, San Mateo, CA: Morgan Kaufmann Publishers.

Shahan, R.W. and Swoyer, C.V. (eds.) (1979) *Essays on the Philosophy of W.V. Quine*, Hassocks: Harvester.

Shallice, T. (1988) *From Neuropsychology to Mental Structure*, Cambridge: Cambridge University Press.

Siskind, J.M. (1996) 'A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings', *Cognition* 61, pp. 39-91.

Skarda, C.A. and Freeman, W.J. (1987) 'How brains make chaos to make sense of the world, *Behavioral and Brain Sciences* 10, pp. 161-95.

Skinner, B.F. et al. (1984) 'Skinner: Canonical Papers', *Behavioural and Brain Sciences* 7, pp. 473-701.

Smart, J.J.C. (1962) 'Sensations and Brain Processes' in Chappell, V.C. (ed.) (1962) *The Philosophy of Mind*, Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Smith, M. (1994a) 'Why Expresivists about Value Should Love Minimalism about

Truth', *Analysis* 54, pp. 1-12.

Smith, M. (1994b) 'Minimalism, Truth-aptitude and Belief', *Analysis* 54, pp. 21-6.

Smith, P. (1981) *Realism and the Progress of Science*, Cambridge: Cambridge University Press.

Smolensky, P. (1988) 'On the Proper Treatment of Connectionism', *Behavioral and Brain Sciences* 11, pp. 1-74.

Smolensky, P. (1990a) 'Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems', *Artificial Intelligence* 46, pp. 159-216.

Smolensky, P. (1990b) 'Representation in Connectionist Networks', *Intellectica* 9-10, pp. 127-65.

Smolensky, P. (1991) 'Connectionism, Constituency and the Language of Thought' in Loewer, B. and Rey, G. (1991) *Meaning in Mind: Fodor and His Critics*, Oxford: Blackwell.

Smolensky, P., LeGendre, G. and Miyata, Y. (1992) 'Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition', Technical Report 92-08, Institute of Cognitive Science, University of Colorado.

Sober, E. (1984) *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. Cambridge, MA.: MIT Press.

Sterelny, K. (1993) 'Refuting Eliminativism on the Cheap?', *Mind and Language* 8, pp. 306-15.

Stern, K. and McClintock, M.K. (1998) 'Regulation of Ovulation by Human Pheromones', *Nature* 392, pp. 126-7.

Stich, S. (1983) *From Folk Psychology to Cognitive Science*, Cambridge, Mass.: MIT Press.

Stich, S. (1990) *The Fragmentation of Reason*, Cambridge, Mass.: MIT Press.

Stich, S. (1991) 'Do True Believers Exist?', *The Aristotelian Society* 65 (suppl.), pp. 229-44.

Stich, S. (1992) 'What is a Theory of Mental Representation?', *Mind* 101, pp. 243-62.

Stich, S. (1996) *Deconstructing the Mind*, Oxford: Oxford University Press.

Stich, S and Warfield, T (1994) *Mental Representation: A Reader*, Oxford: Blackwell.

Stich, S and Warfield, T (1995) 'Reply to Clark and Smolensky: Do Connectionist Minds Have Beliefs?' in Macdonald, C. and Macdonald, G. (eds.) (1995a) *Philosophy of Psychology: Debates on Psychological Explanation*, Vol. 1, Oxford: Blackwell.

Stroud, B. (1989) 'Quine's Physicalism' in Barrett, R. and Gibson, R. (1989) *Perspectives on Quine*, Oxford: Blackwell.

Tani, J. (1998) 'An Interpretation of the 'Self' from the Dynamical Systems Perspective: A Constructivist Approach', *Journal of Consciousness Studies*, 5, pp. 516-42.

Tennant, N. (1987) *Anti-Realism and Logic*, Oxford: Clarendon Press.

Thelen, E. (1995) 'Time-scale Dynamics and the Development of an Embodied Cognition' in Port, R. and van Gelder, T. (1995) *Mind as Motion*, Cambridge, Mass.: MIT Press.

Thelen, E. and Smith, L. (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*, Cambridge, Mass.: MIT Press.

Tienson, J. (1988) 'An Introduction to Connectionism', *Southern Journal of Philosophy* 26, suppl.: pp. 57-84.

Tiffany, E. (1999) 'Semantics San Diego Style', *The Journal of Philosophy.*

Tomberlin, J. (ed.) (1995) *Philosophical Perspectives*, 9: AI, Connectionism and Philosophical Psychology, Ridgeview.

Toulmin, S and Goodfield, J. (1962) *The Architecture of Matter*, Chigago and London: The University of Chicago Press.

Toureztky, D.S. (ed.) (1989) *Advances in Neural Information Processing Systems I*, San Mateo, CA: Morgan Kaufmann Publishers.

Travis, C. (ed.) (1986) *Meaning and Interpretation*, Oxford: Blackwell.

Tye, M. (1991) 'Representation in Pictorialism and Connectionism' in Horgan, T. and Tienson, J. (eds.) (1991) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic Plublishers.

van Gelder, T. (1990) 'Compositionality: A Connectionist Variation on a Classical Theme', *Cognitive Science* 14, pp. 355-84.

van Gelder, T. (1991) 'What is the 'D' in 'PDP'? A Survey of the Concept of Distribution' in Ramsey, W., Stich, S. and Rumelhart, D.E. (eds.) (1991) *Philosophy and Connectionist Theory*, Hillsdale, N.J.: Lawrence Erlbaum.

van Gelder, T. (1992) 'Making Conceptual Space' in Davis, S. (1992) *Connectionism: Theory and Practice*, New York: Oxford University Press.

van Gelder, T. (1995) 'What Might Cognition Be, if not Computation?', *Journal of Philosophy* 92, pp. 345-81.

van Gelder, T. (1998) 'The Dynamical Hypothesis in Cognitive Science', *Behavioral and Brain Sciences* 21, pp. 615-665

van Straaten, Z. (ed.) (1980) *Philosophical Subjects: Essays Presented to P.F. Strawson*, Oxford: Oxford University Press.

Wallace, J. (1977) 'Only in the Context of a Sentence Do Words Have Any Meaning', *Midwest Studies in Philosophy* II, pp. 144-64.

Warner, R. and Szubka, T. (1994) *The Mind-Body Problem*, Oxford: Blackwell.

Way, E.C. (1997) 'Connectionism and Conceptual Structure', *American Behavioral Scientist* 40/6, pp. 729-53.

Weir, A. (1996) 'Ultramaximalist Minimalism!', *Analysis* 56, pp. 10-22.

Wheeler, M. (1994) 'From Activation to Activity: Representation, Computation and the Dynamics of Neural Network Control Systems', *Artificial Intelligence and Simulation of Behaviour Quarterly* 87, pp. 36-42.

Wiggins, D. (1980) 'What Would Be a Substantial Theory of Truth?' in van Straaten, Z. (ed.) (1980) *Philosophical Subjects: Essays Presented to P.F. Strawson*, Oxford: Oxford University Press.

Wilson, N.L. (1958) 'Substance without Substrata', *Review of Metaphysics* 12, pp. 521-39.

Wright, C. (1981) 'Rule-Following, Objectivity and the Theory of Meaning' in Holtzmann, S. and Leich, C. (eds.) (1981) *Wittgenstein: To Follow a Rule*, London: Routledge.

Wright, C. (1983) *Frege's Conception of Numbers As Objects*, Aberdeen: Aberdeen University Press.

Wright, C. (1986a) *Realism, Meaning and Truth*, Oxford: Blackwell.

Wright, C. (1986b) 'Theories of Meaning and Speakers' Knowledge' in Wright (1986a).

Wright, C. (1986c) 'How Can the Theory of Meaning Be a Philosophical Project?', *Mind and Language* 1/1, pp. 31-44.

Wright, C. (1992) *Truth and Objectivity*, Cambridge, Mass.: Harvard University

Press.

Wright, C. (1993) 'Eliminative Materialism: Going Concern or Passing Fancy?', *Mind and Language* 8, pp. 316-26.

Wright, C. (1994) 'Response to Jackson', *Philosophical Books* 35, pp. 169-75.

Wright, C. (1997) 'The Indeterminacy of Translation' in Hale, B. and Wright, C. (eds.) (1997) *A Companion to the Philosophy of Language*, Oxford: Blackwell.