A Connectionist Defence of the Inscrutability Thesis

FRANCISCO CALVO GARZÓN

Abstract: This paper consists of four parts. In section 1, I shall offer a strategy to bypass a counter-example which Gareth Evans (1975) offers against Quine’s Thesis of the Inscrutability of Reference. In section 2, I shall introduce a criterion recently produced by Crispin Wright (1997) in terms of ‘psychological simplicity’ which threatens the perverse route offered in section 1. In section 3, I shall argue that a LOT model of human cognition could motivate Wright’s criterion. In section 4, I shall argue that if we instead model human cognition by a recurrent neural network, then Wright’s criterion is unmotivated. Thus I shall produce a Connectionist Defence of the Inscrutability Thesis.

First let me briefly run through Quine’s argument for the Inscrutability of Reference. 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 (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 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. 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

Versions of this paper were presented at the 1997 University of Glasgow-University of St. Andrews Joint Philosophy Conference at the Isle of Raasay (Scotland), and at the 1998 EPL meetings at the University of California, San Diego. I wish to thank the audiences at these two meetings. I am especially indebted to Jim Edwards for many helpful discussions on these and related topics during the last few years. I am also grateful to Paul Churchland, Jeff Elman, and two anonymous referees for Mind and Language for helpful comments and suggestions. This research was supported by a grant from The Darwin Trust of Edinburgh, and by a grant from the Caja Murcia Savings and Loan Association of Spain (Fulbright Program).

Address for correspondence: Departamento de Filosofía, Edificio Luis Vives, Universidad de Murcia, E-30100 Murcia, Spain.

Email: fcalvo@um.es

© Blackwell Publishers Ltd. 2000, 108 Cowley Road, Oxford, OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.
to be the set of undetached rabbit parts. 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, this hope is thwarted, Quine argues, since the apparatus of individuation is itself inscrutable too.2

1. Semantic Perversity

To introduce Evans’ counter-example, consider two semantic theories of Native, one standard and the other perverse.3 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)
(a1) (x)(x satisfies ‘blanco’ \(f\) iff (x is white & x satisfies \(f\))

Theorem:
(a2) (x)(x satisfies ‘blanco’ \(\land\) ‘gavagai’ iff (x is white & x is a rabbit))4

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

PT
Axioms:
(b) (x)(x satisfies ‘gavagai’ iff x is an undetached rabbit part)
(b1) (x)(x satisfies ‘blanco’ \(f\) iff (x is white & x satisfies \(f\))

Theorem:

---

1 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).

2 The reasons for this are well-known and I shall not pursue them here. The reader not familiar with them may consult Quine, 1960, ch. 2.

3 Although Quine initially employed his parable to illustrate the Indeterminacy of Translation, 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 to 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.

4 (a2) is obviously a consequence of (a) and (a1). 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.
Let us suppose that ST is behaviourally adequate. We can, thus, identify the sentence ‘Blanco gavagai’ with ‘White rabbit’. However, Evans argues, if ST is behaviourally adequate, then PT₁ is not behaviourally adequate. There are certain circumstances in which PT₁ fails to reflect correctly the native’s linguistic behaviour (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, PT₁ is not faithful to the evidence since, assuming PT₁, natives should assent to ‘Blanco gavagai?’ when stimulated by a white-legged brown rabbit.⁵ But, we have assumed that ST is behaviourally correct, and hence that natives would assent to the combined construction ‘Blanco gavagai?’ only in the 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:

\[(b₁)*(x)(x \text{ satisfies ‘blanco’ iff } x \text{ is an undetached part of a white rabbit})\]

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 \((b₁)*\) in order to talk about white things other than rabbits. Hence, the broader axiom required should run as follows:

\[(b₁)**(x)(x \text{ satisfies ‘blanco’ iff } x \text{ is an undetached part of a white thing})\]

But, as Evans notices, the Quinean still faces a similar worry to the one motivated by white-legged brown rabbits. According to \((b₁)**\), ‘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.

However, there is a rejoinder available to the Quinean. I propose a particular way to discriminate among schemes of reference denoting diverse undetached rabbit parts. We can 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 perverse scheme may claim that ‘gavagai’ divides its reference over 5% of the whole rabbit, including 5% of

⁵ Notice that a brown rabbit’s white leg is a white undetached rabbit part.
its surface (henceforth abbreviated 5%-urp:—i.e., 5% undetached rabbit part). Another scheme, over 80%-urp, and so on.6

Let’s see how manuals dealing with x%-urp can cope with Evans’ white-legged brown rabbit. 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 to or bigger than \( \beta \), we take the rabbit to be white. Now, my contention is that a perverse scheme that divides the reference of ‘gavagai’ over \( \beta \)-urp, or greater, will cope with Evans’ white-legged brown rabbit: Take \( \beta \) for instance, as 99. The perverse theory would then run as follows:

\[
\text{PT}_2 \\
\text{Axioms:} \\
(c) \ (x \text{ satisfies ‘gavagai’ } \text{ iff } x \text{ is a } 99\%- \text{ urp}) \\
(c_1)(x)(x \text{ satisfies ‘blanco’}^f \text{ iff } (x \text{ is white } \& \ x \text{ satisfies } f)) \\
\text{Theorem:} \\
(c_2)(x)(x \text{ satisfies ‘blanco’}^\wedge ‘gavagai’ \text{ iff } (x \text{ is white } \& \ x \text{ is a } 99\%- \text{ urp}))
\]

Now, let’s see how this perverse referential scheme behaves under Evans’ pool of data. The question is: Would the native guided by PT\(_2\) 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 white. Hence, Evans’ counter-example is not a counter to PT\(_2\). 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\(_2\) misrepresents native usage.7 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 with respect to the kind of linguistic behaviour considered by Evans.

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 refer-

---

6 Notice that pointing cannot help to solve the referential inscrutability: every time you point to a rabbit, you are pointing to 5%-urps, to 80%-urps, etc.

7 The reader should notice that this result won’t be achieved via any arbitrary value we assign to \( \beta \) in the ‘\( \beta \%- \text{ urp} \)’ perverse scheme. By favouring, for example, a ‘5%-urp’ scheme, Evans’ counter would kick in. Think of a brown rabbit with 5% of its surface coloured white.
ence 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 PT2. Since 99 is smaller than 100, there will always be an undetached part of a whole rabbit which does not belong to any given 99%-urp:—namely, a 1%-urp. Hence the perversity of PT2 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. The indeterminacy, thus, remains unsolved. Native linguistic behaviour, so far as it is considered by Evans, gives us no clue as to whether ‘gavagai’ divides its reference over rabbits or 99%-urp.8

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 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.9 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 Calvo Garzón, under review a, 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 is the only correct semantic theory.

Let us elaborate on Wright’s argument. Wright claims that

\[\text{(A) the basic clauses of our semantic theory are to assign reference and satisfaction-conditions in ways which are presumed to correspond to the conceptual repertoire of speakers of the language in question. (Wright, 1997, p. 412, Wright’s emphasis)}\]

8 The reader may care to consult Hookway, 1988, for a different strategy to bypass Evans’ counter-example. However, as I’ve argued elsewhere (Calvo Garzón, under review a), the ‘99%-urp’ perverse route is not subject to certain criticisms which may put Hookway’s proposal in jeopardy.

9 Wright targets a perverse route offered by Hookway, 1988, in terms of disjunctive axioms, but his argument, if valid, applies equally to PT2.

© Blackwell Publishers Ltd. 2000
It may seem that Wright begs the question against Quine. Obviously, a hard-line 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. Whatever naturalizing strategy we adopt—see sections 3 and 4 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.10 Wright’s argument can then proceed as follows. Firstly, Wright notices, with respect to putative perverse alternatives to the standard scheme, that

\[(B) \text{ the range of concepts necessary in order to formulate their various } \text{[basic] clauses in each case includes, but is not included in, the simple range of concepts of observable spatio-temporal continuants and their observable properties which the favoured scheme deploys.} \text{ (Wright, 1997, p. 412)}\]

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:

\[(C) \text{ To have the concept of an undetached rabbit part, you need a concept of the integrated individual of which such parts are parts; to have the concept of a temporal stage of a rabbit, you need to grasp the idea of the spatio-temporal continuant of which such a stage is a stage.} \text{ (Wright, 1997, p. 412)}\]

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 section 1.11

However, we need to guard against a misreading of the above argument.

---

10 Cf. Quine, 1975, p. 95: ‘To cite a behavioral 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’.

11 An anonymous referee of this journal points out that the ‘psychological complexity’ imputed to the native by a perverse semantic theory might 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. In Calvo Garzón, under review a, I argue that the ‘99%-urp’ referential scheme would conform to such a constraint.
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 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. Suppose we are devising translation manuals for fellow speakers. I may translate your English sentence ‘There is 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 manuals.

Wright reads the basic clauses realistically—as we saw in the quote labelled (A) above. Hence, he takes the conceptual repertoire of the basic clauses to be subject to a manifestation requirement. 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 PT₂, whereas a connectionist setting is neutral between ST and PT₂. The remainder of this paper will be devoted to developing this argument.

3. Psychological Simplicity and Classical Constituency

According to a LOT hypothesis we, as thinkers, have the capacity to entertain thoughts with particular contents which are carried by the mental represen-

---

12 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, Inscrutability of Reference would amount to no more than a trivial—as far as semantics is concerned—clash of cultures.
tations of LOT. For example, to entertain the thought THERE IS A WHITE 99%-URP OVER THERE,\(^{13}\) 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 concepts is the occurrence of word-tokens of LOT.\(^{14}\) 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. (Fodor, 1987, p. 137)

Classical constituency 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. 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.

\(^{13}\) From now on I shall use capital letters to express concepts.

\(^{14}\) See Fodor, 1975; 1987. There are different versions of the LOT hypothesis. Some people maintain that LOT is actually the thinkers’ spoken language, but internalized. Others take LOT to be the analogue 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).

© Blackwell Publishers Ltd. 2000
4. Simple Recurrent Networks and Conceptual Inclusion

Research in neurobiology and cognitive science, however, appears to favour a drastically different picture of what goes on in our heads when mastering natural languages and thinking (see, for example, Churchland and Sejnowski, 1992; and Elman et al., 1996). The processing of natural languages—where information is encoded serially—calls for the representation of complex hierarchical grammatical structures. A connectionist network\(^{15}\) 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.\(^{16}\)

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 recurrency the network can process contextualized sequential information. Based on this recurrent architecture, Elman (1992)—see Figure 1—designed a network which does exhibit appropriate sensitivity to the syntactical dependencies found in grammatical structures. 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

\(^{15}\) I shall assume familiarity with the basic tenets of connectionism. For an introduction, the reader may care to consult Bechtel and Abrahamsen, 1991.

\(^{16}\) 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 architecture involves a connectionist model of cognition. Nevertheless, I hope that by the end of the paper an idea of how to answer the charge of ‘mere implementation’ will begin to emerge. 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.

© Blackwell Publishers Ltd. 2000
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 discriminate grammatically acceptable structures from grammatically unacceptable ones. Being fed with a sequence of words from the input stream, the network had to predict the subsequent word. Using backpropagation—see Rumelhart, Hinton and Williams, 1986—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. Obviously, the prediction task for the net cannot be deterministic. Given a novel input, several correct outputs may follow. 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. 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.

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’, and  
(b) ‘boys, who boys chase, chase boy’.

Figure 1 Elman’s recurrent network used to discriminate grammatically correct sentences. (from Elman, 1992, p. 153. Used by permission).
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. Elman’s net respected the grammatical agreement between the main clause subject and the main clause verb. 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 (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 behaviour—we can see the temporal trajectories in the processing of sentences. 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 Figure 2—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. The representations obtained in Elman’s net are fully distributed. 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 unstructured representations. Grammatical structure is reflected by coding grammatical vari-

---

17 Similarly, Elman’s net could represent successive embedding relationships, as found in complex relative clauses. See Elman, 1992, pp. 165–7, for the details.

18 For other philosophical implications of Elman’s net see Ramsey, 1992, pp. 269–71.
Figure 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. Used by permission).

ations 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. Connectionist constituency is *context-dependent*. 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, I contend, we find no motivation for Wright’s argument.19 Imagine we feed Elman’s recurrent network with several ‘rabbit’-related sentences.20 Take the following sentences:

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

---

19 An anonymous referee of this journal has suggested that coding grammatical variations as slightly different positions in the representational hidden space of Elman’s model does not imply that the model is not subject to a *symbolic* treatment. That is, were we to do PCA on a classical model, we would get results 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, we might obtain context-independent representations in Elman’s model by paying attention to particular principal components, closing thus the gap between the connectionist and the classical approaches. I think this line of argument runs into either of two problems: On the one hand, if one permits infinite precision analog, then indeed it is not clear that their representational capacities can be distinguished from the representational capacities of a recurrent network. However, such a move gives up precisely what makes symbols attractive, namely, their ability to abstract over classes of items, ignoring context. On the other hand, using PCA analysis to ‘symbolicize’ the hidden unit representations might give the appearance that they are equivalent to symbols; but this fails 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 subtle variations in hidden unit representations may have causal consequences for the network’s behaviour—Cf. Servan-Schreiber, Cleeremans and McClelland, 1989. Space, however, prohibits me from elaborating on these matters.

20 This is just a thought-experiment. I shall ignore the technical adjustments required in the design and training of the net.
(c) and (d) are composed out of a set of lexically simple items: Namely, ‘White’, ‘rabbit’, ‘99%’, ‘undetached’ and ‘part’. 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 role₁], ‘rabbit’-qua-[syntactic/conceptual role₂], where syntactic/conceptual role₁ is replacing f in ‘White f’ and syntactic/conceptual role₂ 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 common to ‘rabbit’-qua-[syntactic/conceptual role₁] and ‘rabbit’-qua-[syntactic/conceptual role₂]. 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 role₁] and ‘rabbit’-qua-[syntactic/conceptual role₂] 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 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.\textsuperscript{21}

I conclude then that the Quinean can go with modern scientific fashion and make use of the ‘99%-urp’ referential scheme. By approaching the issue of the naturalization of concepts from a LOT perspective, I argued, Wright may hold to his principle of psychological simplicity. On the other hand, if we identify concepts with non-classical features acceptable to a connectionist, then, as we’ve seen, Wright’s argument does not go through.\textsuperscript{22,23}

Department of Philosophy
University of Glasgow

Department of Philosophy
University of California, San Diego

References


Calvo Garzón, F. (under review b): Is simplicity alethic for semantic theories?


Elman, J. 1998: Generalization, simple recurrent networks, and the emergence of struct-

\textsuperscript{21} It is worth remarking that on the connectionist view the finite set of basic clauses of a translational manual do 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.

\textsuperscript{22} Wright, 1997, produces another argument against the Inscrutability Thesis based on \textit{structural}, rather than psychological, simplicity. The reader may care to consult Calvo Garzón, under review b, for a rejoinder to Wright’s other argument.

\textsuperscript{23} Paul Churchland (personal communication) worries that a connectionist sympathizer of Wright may appeal to a ‘mathematical measure’ of \textit{conceptual similarity}—see Laakso and Cottrell, 1998; and Churchland 1998—in order to achieve a one-to-one translational mapping which favours a standard interpretation of Native. In Calvo Garzón, 2000, I argue 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 neural networks.


© Blackwell Publishers Ltd. 2000