pen-ultimate draft

State Space Semantics

and

Conceptual Similarity:

Reply to Churchland

FRANCISCO CALVO GARZÓN
Visiting Researcher
Department of Philosophy, 0119
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093

Email: fcalvoga@ucsd.edu
State Space Semantics and Conceptual Similarity: Reply to Churchland

**Abstract:** Jerry Fodor and Ernest Lepore (1992; 1996) have launched a powerful attack against Paul Churchland’s connectionist theory of semantics—aka *State Space Semantics*. 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; forthcoming) 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. In this paper 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.
State Space Semantics and Conceptual Similarity: Reply to Churchland

I STATE SPACE SEMANTICS: THE PROBLEM

Paul Churchland (1986) proposed a new, connectionist-inspired, approach to the theory of mental representation known as 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)

Briefly, 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.

1 I shall assume familiarity with the basic tenets of connectionism. For a philosophy-oriented introduction, the reader may care to consult Bechtel and Abrahamsen (1991).
Jerry Fodor and Ernest Lepore (1992; 1996) have recently launched a powerful attack against Churchland’s proposal. One of their objections can be summarized as follows:\(^2\) State Space Semantics understands conceptual similarity across networks as similarity in the activation patterns across those dimensions that specify the networks’ representational spaces.\(^3\) 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 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

---

\(^2\) 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 (forthcoming). For a rebuttal of Churchland’s general strategy to bypass Fodor and Lepore’s criticism see Calvo Garzón (in preparation b).

\(^3\) 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.
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.⁴ (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 II 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).

Before getting started, let me briefly outline the logical form of the paper. In section II, I shall introduce a mathematical technique for measuring conceptual similarity that Aarre Laakso and Gary Cottrell have recently offered in order to address Fodor and Lepore’s challenge. In section III, 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 construe a robust theory of semantics, and a robust theory of translation. In section IV, 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.

⁴ 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.
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. Conclusions and suggested directions for future research will follow in section V.

II A CONNECTIONIST MEASURE OF CONCEPTUAL SIMILARITY

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

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 in section I, 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 similar—see 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.

Laakso and Cottrell (1998; forthcoming) have recently taken up Fodor and Lepore’s challenge. 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 (forthcoming). 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:

\[
A = <0>, \quad B = <50>, \quad \text{and} \quad C = <100>.
\]

On the other hand, network #2 represents the same three things with the following vectors:

\[
A = <0, 0>, \quad B = <30, 30>, \quad \text{and} \quad C = <80, 0>.
\]
We can then form the following matrices (see fig. 1) 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. In our example, the two vectors are:

\[<50, 100, 50>, \text{ and } <42, 80, 58>\]

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. 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

5 For the details of both experiments, see Laakso and Cottrell (1998).
Semantics. The question I would like to pursue next is to what extent 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 the paper 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.

III 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. Jeff Elman (1992)—see fig. 2—designed a recurrent network which exhibits appropriate sensitivity to the syntactical dependencies found in sentences.

6 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 I. Both Laakso and Cottrell (1998), and Churchland (1998) seem, at times, to be addressing Fodor and Lepore’s challenge in its epistemic line. 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). An anonymous referee for Philosophical Psychology urges that Churchland’s epistemic reading may be evading the real issue prompted by 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 is rooted in different grounds.
A simple recurrent network is a standard feedforward net supplemented with one or more feedbackward pathways. This recurrent architecture allows the network to deploy 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.

Elman’s network was trained on a set of sentences produced out of a lexicon of 8 nouns, 12 verbs, the relative pronoun ‘who’ and an end-of-sentence period. Being fed with a sequence of words from the input stream, the network’s task was to predict the subsequent word. Using backpropagation weights were adjusted to the desired output performance. Probabilities of occurrence for all possibly correct predictions were determined by generating likelihood vectors for every next word in the novel corpus of sentences. The results Elman reported suggest that the network could successfully discriminate grammatical strings of words—for the data, see Elman (1992).^7

---

^7 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 (thanks to an anonymous referee of Philosophical Psychology for bringing this potential source of misunderstanding to my attention). We may say, however, that the ungrammaticality of "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
An example will help to illustrate 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.

The results of the prediction task 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. Making conceptual sense of the processing is not straightforward and requires some simplifying statistical treatment. We need to observe the temporal trajectories of the hidden patterns through state space. Principal Components Analysis (PCA) provides us 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 judgment’. 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.

8 Note that number information (e.g., boy/s) needs to be taken into account over the relative clause—who boys chase—common to (a) and (b) in order to predict the final “chases” or “chase”. Similarly, Elman’s net could represent successive embedding relationships, as found in complex relative clauses. See Elman, 1992, pp. 165-7, for the details.
with a relatively simple way of looking into this high-dimensional sequential vector space. PCA is a technique for reducing the number of dimensions which consists in passing each member—sentence—of the input set used in the training 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 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. 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. 3—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.

[INSERT FIG. 3 AROUND HERE]

Churchland (1998) considers how Laakso and Cottrell’s experiments might apply to the case of simple recurrent networks. As figure 3 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, 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)

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)

I shall argue next that Churchland’s defence of State Space Semantics fails to bring robustness to semantic discourse. In particular, I shall argue that Churchland lacks a

9 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. 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.
connectionist notion of synonymy of the kind required by a robust theory of translation, and by extension, by a robust theory of mental representation.

IV A CONNECTIONIST APPROACH TO RADICAL TRANSLATION: REPLY TO CHURCHLAND

For argument’s sake, I shall 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.\(^\text{10}\) 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 section V 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

\(^{10}\) 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 individuals are linked to the world in similar ways or not. Fleshing out this thought would take us far afield from our present purposes.
connectionist reading of Quine’s Thesis of the Inscrutability of Reference, in order to argue that State Space Semantics cannot provide such a robust criterion.¹¹

Let me review very briefly the main contention of Quine’s Inscrutability Thesis. Take, for instance, the Native sentence ‘Blanco gavagai’. It’s been empirically determined that ‘Blanco gavagai’ relates to portions of space-time, in the vicinity of the native speaker, which are white-related and rabbit-related. We may then translate ‘Blanco gavagai’ with our ‘Lo, a white rabbit’. However, it would be rash to impute our ontology to the natives. Quine maintains that the extension of the term ‘gavagai’ could be taken to be the set of undetached rabbit parts. His conclusion is that it is inscrutable what the expression ‘gavagai’ refers to: The semanticist may assign to the native expression as its extension either the set of rabbits or the set of undetached rabbit parts.¹² Imagine then three semanticists, each devising a semantic theory in order to produce behaviourally supported truth conditions for any native sentence. Each semanticist comes out with a different semantic theory. They are different in the non-trivial sense that they assign different extensions to some of the native terms of the language under study.

¹¹ 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. Focusing on the theory of reference from a Quinean perspective simply shows my personal biases. For an overall attack to Churchland’s general theory of content see Calvo Garzón (in preparation b).

¹² The reader not familiar with Quine’s arguments may care to consult his Word and Object, chapter 2. 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.
Take, for example, the Native compound expressions ‘blanco gavagai’ and ‘blanco gato’. A Standard Theory, ST, would deal with the satisfaction conditions of those expressions in the following way:

**ST**

Axioms:

(a) \( (x)(x \text{ satisfies ‘gavagai’ iff } x \text{ is a rabbit}) \)

(a₁) \( (x)(x \text{ satisfies ‘gato’ iff } x \text{ is a cat}) \)

(a₂) \( (x)(x \text{ satisfies ‘blanco’} ^f \text{ iff (} x \text{ is white & } x \text{ satisfies } f) \)

Theorems:

(a₃) \( (x)(x \text{ satisfies ‘blanco’} ^‘gavagai’ \text{ iff (} x \text{ is white & } x \text{ is a rabbit}) \)

(a₄) \( (x)(x \text{ satisfies ‘blanco’} ^‘gato’ \text{ iff (} x \text{ is white & } x \text{ is a cat}) \)

On the other hand, a Hybrid Theory—i.e., standard-cum-perverse—, HT, would run as follows:

**HT**

Axioms:

(b) \( (x)(x \text{ satisfies ‘gavagai’ iff } x \text{ is an undetached rabbit part}) \)

(b₁) \( (x)(x \text{ satisfies ‘gato’ iff } x \text{ is a cat}) \)

(b₂) \( (x)(x \text{ satisfies ‘blanco’ iff either} \)

---

13 (a₃) and (a₄) are obviously a consequence of (a), (a₁), and (a₂). 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.
(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:

(b3) (x)(x satisfies ‘blanco’ \( \land \) gavagai iff (x is an undetached part of a white rabbit))

(b4) (x)(x satisfies ‘gato’ \( \land \) ‘blanco’ iff (x is white \& x is a cat))

And finally, a fully Perverse Theory, PT, would account for the Native compounds in the following way:

PT

Axioms:

(c) (x)(x satisfies ‘gavagai’ iff x is a 99% undetached rabbit part)\(^{14}\)

(c1) (x)(x satisfies ‘gato’ iff x is a 99% undetached cat part)

(c2) (x)(x satisfies ‘blanco’ \( \land \) f iff (x is white \& x satisfies f))

Theorems:

(c3) (x)(x satisfies ‘blanco’ \( \land \) ‘gavagai’ iff (x is white \& x is a 99% undetached rabbit part))

\(^{14}\) We may talk in terms of the percentage of the whole rabbit, including the percentage of its surface, that we assign as the extension of ‘gavagai’. In this way, axiom (c) says that ‘gavagai’ divides its reference over an undetached 99% part of the whole rabbit, including 99% of its surface, and similarly for ‘gato’ in (c1)—see fn. 15 below.
(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, HT and PT are behaviourally fully adequate too. A translator guided by either HT or PT 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.\(^\text{15}\)

Imagine now that we train a simple recurrent network, call it N, on Native sentences of which these are examples:

\[
\begin{align*}
\text{(1)'} & \quad \text{Blanco gavagai.} \\
\text{(2)'} & \quad \text{Blanco gato.}
\end{align*}
\]

Also we train three simple recurrent networks—call them network A, network B and network C—with English sentences derived from ST, HT and PT, respectively.\(^\text{16}\)

Sentences for network A are:

\[
\text{(I) There is a white rabbit.}
\]

\(^{15}\) The reader may wonder why a perverse semanticist would employ HT or PT. HT relates to a sketched theory of translation offered by Christopher Hookway (1988) in order to bypass a counter-example that Gareth Evans (1975) offered against Quine’s Inscrutability Thesis. PT corresponds to a different strategy to bypass Evans’ counter which I have developed elsewhere (Calvo Garzón, forthcoming; under review a).

However, for our purposes, we need only bear in mind that HT and PT are empirically adequate whenever ST is empirically adequate. The reasons for using both HT and PT will become apparent in a moment.

\(^{16}\) 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.
(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 III), State Space Semantics should furnish us with an objective criterion for judging sameness of content which will deliver an accurate translational map.¹⁷ Imagine then that we apply Principal Components

¹⁷ An anonymous referee for Philosophical Psychology points out that the truth-conditional semantics invoked to spell out ST, HT, and PT might be at odds with Churchland’s connectionist approach to semantics. This, however, should not cause any concern. We may naturalize concepts, 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 would agree—ST, HT, and PT 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 \(<h_1, \ldots, h_n>\) across the hidden units \(\{H_1, \ldots, H_n\}\) will carry information about
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, 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 ‘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 contrasting N with B, and with C—fully perverse English? Churchland certainly does not want this to be the 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 HT, and PT will deliver representations identifiable, along other dimensions—along UNDETACHED/RABBIT/PART, and 99%/UNDETACHED/RABBIT/PART semantic dimensions, respectively.
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, HT), Churchland may be right, in the case of C (our fully perverse theory, PT), 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 HT, the satisfaction conditions of ‘blanco’ are linked to undetached parts of white—... when ‘blanco’ is coupled with ‘gavagai’. In all other cases, HT 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. 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 with respect to the trajectories obtained in the Native Network.18 Nevertheless, 18 The reader may wonder whether we could broaden the scope of HT’s perversity. Axiom (b₂) in HT would then need to 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, paper, etc., etc. And the same would happen with respect to all those axioms required for dealing with any other Native colour-word—see Calvo Garzón (under review a). 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. 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 HT, we would have a perfect isomorphism with respect to Native, which is what I aim to show now with PT.
the hybrid character of HT (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 achieves.

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.\textsuperscript{19} 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 shows, I contend, 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)’—or, at least, no grounds in the light of Churchland’s considerations.\textsuperscript{20}

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

\textsuperscript{19} Note that derivations in PT have exactly the same syntactic structure as derivations in the standard theory, ST.

\textsuperscript{20} Someone may contend 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. For arguments against structural and psychological simplicity constraints, see Calvo Garzón (forthcoming; under review b), respectively.
of content.\textsuperscript{21} As I mentioned above, Churchland’s way of determining content comes
\textit{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 II). We need to consider the \textit{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 \textit{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. 11 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 that there is \textit{no fact of the
matter} as to which objective macrofeatures are the ones being pinned down.\textsuperscript{22}
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, \textit{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)

\textsuperscript{21} Thanks to an anonymous referee of \textit{Philosophical Psychology} for bringing this worry to my attention.

\textsuperscript{22} For a connectionist defence of Quine’s Inscrutability Thesis see Calvo Garzón (forthcoming).
I ignore what moves Churchland to make such a strong contention.\textsuperscript{23} 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 \textit{observationally indistinguishable}, 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, \textit{contra} Churchland, as a consequence of sharing a common conceptual framework. But let us move on now 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 \textit{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 \textit{fine-grained} structure that hopefully can be distinguished under principal components analysis.\textsuperscript{24} Someone sympathetic to Churchland may then hope to exploit

\textsuperscript{23} Indeed, “[...] input-output behaviors are nevertheless identical, \textit{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).

\textsuperscript{24} 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
these potential fine discriminations in the following way: If network C is trained to achieve grammatical competence on a large set of fully-perverse sentences, then it will have to master, among other things, the grammar of \textit{percentile fractions}, the grammar of \textit{wholes} and \textit{parts}, both \textit{detached} and \textit{undetached}, and a substantial \textit{vocabulary} that is \textit{absent in the coding activity of network A}. We may therefore be able to discriminate between the two networks by examining their 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 \textit{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 \textit{no} vocabulary deployed by network C that is absent in the coding activity of network A. We can see this by devising translation manuals for the fellow speakers of our Home language.\textsuperscript{25} 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 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.

\textsuperscript{25} Note that setting the parable of Radical Translation in a \textit{home} environment—i.e., English-to-English translation—should not alter matters significantly. The success of the Inscrutability Thesis cannot be dependent on the object-language being \textit{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.
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 need to translate it and can do so either via my standard manual or via my perverse manual. Obviously the process iterates indefinitely. This neatly illustrates the fact that whatever vocabulary network C deploys will also be present in the coding activity of network A. In short, the toy-languages that networks A and C are trained on are formed out of the same lexicon. We won’t then be able to discriminate between then 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—see fn. 12 above—, 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. 31.26

V CONCLUSION

26 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, HT, and PT, respectively, and computing the correlations of trajectories across networks. I hope that these neurosimulations will back up the theoretical argument of this section.
In this paper, I have argued that appealing to Laakso and Cottrell’s mathematical measure of conceptual similarity cannot bring Churchland’s optimistic conclusion. Namely, the conclusion that State Space Semantics can furnish us with a robust 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; forthcoming) 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 paper 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.27

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 paper might have a broader impact than I have argued for here. For strategical reasons, I’ve assumed throughout this paper 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:

27 It is no surprise to me that Quine would agree with a connectionist setting of his ‘radical translation’ arguments. Sometime before the explosion of connectionist works in the 80s he comments: *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) Well, the time has come to submit Quine’s arguments to a neurocomputational explanation (see Calvo Garzón, forthcoming).
A LOT model with classical constituency, and classical processing, or a connectionist model where constituency and processing are non-classical—see Calvo Garzón, forthcoming. 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.28, 29

References


28 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.

29 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, Gary Cottrell, Jeff Elman, Aarre Laakso, and two anonymous referees for Philosophical Psychology for helpful comments and suggestions. This research was supported by a grant from the Caja Murcia Savings and Loan Association of Spain (Fulbright Program).
CALVO GARZÓN, F. (under review b). Is Simplicity Alethic for Semantic Theories?


State Space Semantics and Conceptual Similarity: Reply to Churchland

Francisco Calvo Garzón

Fig. 1
State Space Semantics and Conceptual Similarity: Reply to Churchland

Francisco Calvo Garzón

Fig. 2
State Space Semantics and Conceptual Similarity: Reply to Churchland

Francisco Calvo Garzón

Fig. 3
List of captions for figures 1-3

[Fig. 1]: Symmetric matrices obtained by taking Euclidean distances between all the representations in each network. (From Laakso and Cottrell, under review)

[Fig. 2]: Elman’s recurrent network used to discriminate grammatically correct sentences. (From Elman, 1992, p. 153)

[Fig. 3]: 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)