Version of 2/12/99

All AT SEA IN SEMANTIC SPACE: CHURCHLAND ON MEANING SIMILARITY

Jerry Fodor and Ernie Lepore
Center for Cognitive Science
Rutgers University

Introduction

It's old news that the identity conditions for contents, meanings and the like are problematic, perhaps because there can't be identity conditions for meanings unless there is an analytic/synthetic (a/s) distinction, and there isn't. We're not, in the present paper, proposing to consider whether the notion of content identity can be made metaphysically respectable. But we want to discuss a proposal about how to save semantics if it does turn out that 'same meaning' can't be reconstructed; it's one that is currently popular in both the philosophical and the cognitive science literature.[1]

The idea is that the purposes semantic theories are supposed to serve don't actually require principles of content *individuation;* all they require is a notion of content *similarity.* On such a view, similarity of meaning is primitive relative to sameness of meaning, and the latter might be of interest only as a limiting case of the former; perhaps, indeed, a limiting case that is never achieved in practice. Meaning similarity is real and robust, there's lots of it around, and it's metaphysically prior to synonymy. So the story goes.

We will have two points to make. First, even if a notion of content similarity were on offer, it would arguably be of no use for doing the kinds of things that a theory of meaning ought to do. *Quite generally, the kinds of explanations that semantic theories are supposed to give would not survive substituting a similarity-based notion of content for an identity-based notion.* This is unaccidental; it's the result of ways in which content similarity and content identity are intrinsically different kinds of relations. Second, we'll argue that, even if a similarity based notion of content could in principle replace an identity based notion *salve* the explanatory power of semantic theories, that wouldn't be much cause for rejoicing. For, as a matter of fact, nobody has the slightest idea how to construct the required notion of content similarity; and the reasons for this are much the

[1] Texts where this sort of suggestion is endorsed are ubiquitous. For some examples: Harman, G., 1973, Thought, Princeton: Princeton University Press; Block, Ned, "Advertisement for a Semantics for Psychology", Midwest Studies in Philosophy, vol. 10, Studies in the Philosophy of Mind, eds. P. French, T. Uehling, H. Wettstein, University of Minnesota Press, Mpls., MN, pp. 615-78, 1986; Sloman, S. and L.J. Rips, Similarity and Symbols in Human Thinking, MIT Press, Cambridge, Mass., 1998; Stich, S, "On the ascription of Content," ed., A. Woodfield, Thought and Object, Clarendon Press, Oxford, pp.153-206, 1981; Jackendoff, Ray, "Conceptual Semantics," eds. Eco, U, M. Santambroggio, P. Violi, Meaning and Meaning Representations, Indiana Press, Bloomington, IN, pp. 81-97, 1988.

same, *mutatis mutandis*, as the reasons why nobody has the slightest idea how to construct a perspicuous notion of content identity. Until recently, we would have thought this claim was untendentious. But Paul Churchland, in a number of articles over the last decade, has proposed what he says is a schema for such a construction. That he is quite wrong to say this has been the burden of several papers of ours, to which Churchland has now likewise replied.[2] Our view is that his reply is an egregious *ignoratio elenchi*. The second part of the present discussion is our attempt to make clear why it is.

PART 1 – Why content similarity is probably no use.

A question that is not asked nearly as often as it should be is this: If you had a tenable notion of similarity of meaning, *what exactly would you do with it?* We suspect that friends of similarity-based content more or less take for granted that, given such a notion, they could just plug it in wherever semantic theories have hitherto availed themselves of synonymy, content identity, intentional equivalence and the like. It is (we continue to suspect) because the friends of meaning similarity assume this that they rarely bother to show how a semantics built around content similarity could do anything like the explanatory work that semantic theories are supposed to do.[3] In fact, a plausible *prima facie* case can be made against this assumption; so, anyhow, we're about to argue.

We can think of four main purposes for which one might want a theory of meaning and for which such theories have historically been employed. We don't deny that there may be other uses they could be put to: but we do think it's plausible that if a theory of meaning failed all of the following, it wouldn't be worth having.

2.1 Satisfaction Conditions. A semantic theory for the language T ought to assign satisfaction conditions to the semantically evaluable expressions of T; likewise, *mutatis mutandis*, for a theory of the intentional contents of semantically evaluable mental states.[4] Moreover, it should turn out that sentences that we're all quite certain are true *are* true under the preferred assignments of satisfaction conditions. For example, *ceteris paribus* there would be something badly wrong with a semantic theory that assigned truth

[2] Churchland, P. M., "Some reductive strategies in cognitive neurobiology", in A Neurocomputational Perspective: The nature of mind and the structure of science, MIT Press, Cambridge, Mass., pp. 77-110, 1991; Fodor, J. and E. Lepore, Chapter 6 of Holism: A shopper's guide, Basil Blackwell, Oxford, 1992; Churchland, P.M., "Fodor and Lepore: State-space semantics and meaning holism", Philosophy and Phenomenological Research, 53, pp. 667-72, 1993; Fodor, J. and E. Lepore, "Reply to Churchland", Philosophy and Phenomenological Research, 53, pp.679-83, 1993; Churchland, P.M., "Second Reply to Fodor and Lepore, in The Churchlands and Their Critics, ed. R. McCauley, Basil Blackwell, Oxford, pp.278-83, 1996; Churchland, P.M., "Conceptual Similarity across Sensory and Neural Diversity: the Fodor-Lepore Challenge Answered", The Journal of Philosophy, 95, no.1, pp.5-32, 1998.

[3] Arguing for this assumption is a burden which semantic *eliminativists*, unlike the friends of content similarity, are of course not require to discharge. There are people around, including Churchland himself in some of his moods, who apparently can't decide whether their metaphysical bottom line is Realism about content similarity or skepticism about meaning. Our discussion is only concerned with the former doctrine.

[4] We assume that linguistic expressions and propositional attitudes are both semantically evaluable; and we are neutral as to which, if either, is ontologically prior. We will generally go back and forth between the semantics of languages and the semantics of mental states as expository convenience suggests.

conditions to ‘Sparrows are birds’, ‘Water is wet’, ‘George Washington was president’, etc. according to which any of them express falsehoods.

2.2 *Compositionality*. A semantic theory should make clear how the satisfaction conditions for complex linguistic expressions are determined by the satisfaction conditions of their grammatical parts; thereby explaining why natural languages are typically productive and systematic. Likewise, a semantics for propositional attitudes should make clear how the satisfaction conditions for beliefs, desires and the like are determined by those of their constituent concepts, thereby explaining why minds are typically productive and systematic.

2.3 *Translation.* Pre-theoretic intuition has it that meaning is what good translations preserve. A semantic theory should provide a notion of meaning according to which this turns out true.

2.4. *Intentional Explanation*. A semantic theory should reconstruct a notion of content that is adequate to the purposes of intentional (e.g., belief/desire) explanation.

We’re about to argue that, quite likely, a similarity-based theory of meaning wouldn’t do any of these four kinds of things; not even if one assumes, *qua* semantic Realist, that an identity based theory of meaning would do the work if only the individuation problem for contents could be solved.

2.1 *Satisfaction conditions*.

Consider, to begin with, any statement (or sentence or thought) you like that you’re prepared to believe is atomic, has the logical form *Fa*, and is true. Nothing much except ease of exposition turns on the how notions like ‘logical form’ and ‘atomic statement’ are deployed; you choose. (In fact, it doesn’t even matter all that much how you construe the notion ‘true’; we’ll return to this.) Take, for example, the sentence (S1). By our lights,

S1. Nixon is dead.

(i) (S1) is certainly true.
(ii) ‘Nixon is dead(at t)’ is true iff Nixon is dead (at t).[5]
(iii) No semantic theory (for English) could be adequate unless it entails (ii) and is consistent with (i).

Query: what would a content-similarity based semantics say about the truth conditions of (S1)?

Well, we don’t know because, of course, there aren’t really any such theories on offer. But we guess the story would go something like this. Think how an old fashioned, identity based semantic theory goes about entailing (ii). Roughly, it assigns the individual

[5] From here on, we ignore issues about context sensitivity; in particular, we won’t worry about implicit indexicals.

Nixon to the singular term 'Nixon'; it assigns the property of *being dead* to the expression 'is dead', and it interprets sentences of the logical form *Fa* to mean that the individual designated by 'a' has the property expressed by 'F'. Well, likewise, *mutatis mutandis,* according to the new dispensation except that, for example, we assign to the expression 'is dead' not the property of being dead, but *a range of properties all of which are similar to being dead,* and we assign to 'Nixon' *a range of individuals all of whom are similar to Nixon*.[6]

But clearly, to proceed in this fashion is simply to give up the hope of assigning satisfaction conditions to formulas in a way that makes, e.g., sentences like (ii) come out true. For example, we suppose that being comatose on one's death bed is pretty similar to being dead; but it's not the case that if Nixon is comatose on his death bed at t, then Nixon is dead at t. If you don't believe us, ask Doctor Kevorkian.

We take it to be just self evident that properties constructed out of semantic similarity won't do what notions like 'expresses', 'denotes', 'means', etc., are supposed to do in statements like ''Nixon' denotes Nixon'; ''is dead' expresses the property of being dead,' ''is dead' means *is dead'* and the like. The problem is simple enough: if two expressions denote, express, or mean the same, then they are coextensive; but if two expressions only denote, express, or mean *something similar*, they needn't be (and typically aren't). And, it's at the heart of the standard way of assigning satisfaction conditions, that what replaces 'F' on the left hand side of instances of the schema '*Fa* is true iff a is F' must be at least coextensive with what replaces it on the right hand side. The moral seems to be that part of the price of switching from a content-identity based semantics to a content-similarity based semantics is giving up the idea that theories of meaning specify the pre-theoretically correct conditions of semantic evaluation of thoughts, sentences and the like.[7]

---

[6] Perhaps, however, names express individual concepts, in which case 'Nixon' is assigned some property similar to the one that 'Nixon' expresses. We're not at all clear what a similarity based semantics should say about singular terms. Since the options for predicates seem a bit clearer, from here on we'll generally run the discussion on them.

[7] We are assuming that the project is to replace a semantics based on intensional identity with one based on intensional similarity; but we're also assuming that the usual notion of identity of *extension* survives. That is, it's not part of the story that the revisionist semantics would appeal to some notion of similarity *of extension* where traditional semantics adverts to extensional identity. (As far as we know, none of the proposals for a similarity based reconstruction of the notion of intension on offer in the literature do call for a similarity based reconstruction of the notion of extension: Either the issue isn't discussed, or the conservative notion of extension is assumed. See, for example, Putnam, Hilary, 1975,'The Meaning of 'Meaning',' in Mind, Language, and Reality, Cambridge: Cambridge University Press.)

In fact, we can't imagine how a similarity-based notion of extension would go. The problem is that the similarity of two extensions would presumably have to be grounded not in their *degree of overlap,* but in the *similarity of the individuals* that belong to their extensions, and who knows how this notion is to be explicated? Thus, for example, the extension of 'male college senior' is, presumably, "more similar" to the extension of 'female college senior' than it is to the extension of 'rock' or of 'prime number'. But none of these sets overlaps at all. By contrast, the set {male truck drivers} is less similar to the set {male college students} than it is to the set {female college students}, though the first and second probably overlap and the second and third do not.

A last consideration under this general head: You might suppose that you could use a similarity based semantics to assign satisfaction conditions in something like the traditional way if you are prepared give up an absolute notion of *truth* in favor of some correspondingly approximate notion. On this view, for example, the right truth sentence for (S1) might be not (ii) but (iv).

iv. 'Nixon is dead' is *similar to being true* (i.e., it's roughly true) iff Nixon is similar to being dead.

We don't actually know of anybody into meaning similarity who clearly endorses this, nor are we sure what, exactly, such a proposal would amount to. But whatever (iv) means, there are surely lots of counter-instances to the schema *formulas of the form Fa are roughly true iff a is similar to being F*. The basic problem is presumably that `similar to', unlike `roughly true' is, of necessity, relativized to respects. Thus, it's true (we suppose) that dogs are similar to cats; but it isn't even roughly true that dogs are cats; or that my cat is a dog,…etc. The equivalence also fails in the other direction. It's roughly true that you can't be the Pope if you were born in the Bronx. But being the Pope and not being born in the Bronx aren't at all the same sort of thing. Both of the present authors are instances to the contrary.

2.2 *Compositionality*.

Compositionality is the idea that, in the typical case, when syntactically complex expressions (/thoughts) are semantically evaluable, their satisfaction conditions are determined by the satisfaction conditions of their syntactic constituents. And it's generally agreed that the crucial condition that must be satisfied in order that compositionality should hold for a class of expressions is that the satisfaction conditions of their constituents should be *context independent*. So, the fact that 'brown dog', 'green dog', 'brown cat', and 'yellow cat' are all compositional in English is part and parcel of the facts that: 'brown' means the same in the environment '__dog' that it does in the environment '__cat'; 'dog' means the same in the environment 'brown…' that it does in the environment 'green…' Etc. Likewise, *mutatis mutandis*, for 'brown Poodle dog' and the like. That compositionality requires context independence is, to repeat, the consensus view, and we simply take it for granted in what follows.

Now the claim that the meaning of a constituent is context independent is essentially the claim that the following schema is valid (S2).[8]

We will, however, briefly consider a case of a similarity based semantic theory that proposes to dispense not only with identity of meaning but also with identity of truth-value. See immediately below.

S2. If m is part of the meaning of 'a' and 'a' is a constituent of 'b', then m is part of the meaning of 'b'.

The point of present interest, however, is that (S3) – which is what you get if you substitute 'is similar to part of the meaning of' for 'is part of the meaning of' throughout (S2) – is surely false.

S3. If m is similar to part of the meaning of 'a' and 'a' is a constituent of 'b', then m is similar to part of the meaning of 'b'.

So, presumably, the meaning of 'cat' is similar to part of the meaning of 'the leopard is on the mat' (cats and leopards are both felines) and part of the meaning of 'the leopard is on the mat' is similar to part of the meaning of 'the explosion caused a lot of damage' (leopards and explosions are both dangerous.) But the meaning of 'cat' is not similar to any part of the meaning of 'the explosion caused a lot of damage'.

The main argument is, we think, simple and obvious: The standard explication of the compositionality of content depends crucially on assuming that the semantic property of a constituent that is preserved when it's embedded in its host is content identity, not just content similarity.

2.3 *Translation.*

The relation *translates*, like the relation 'is similar to the meaning of', is plausibly intransitive. For this reason, lots of friends of content similarity think that, whatever its flaws may be, it's at least likely to be useful in the theory of translation. And, indeed, there are some who think this is the *only* work a semantic theory can legitimately be asked to do because facts about translation are the only *bona fide* semantic facts there are.

We think, however, that this argument is extremely dubious, at least for the case of infinite languages (*viz.*, languages of infinite expressive power, which we suppose to be the only case worth caring about). Suppose that a translation theory between L_1 and L_2 is a finitely specifiable, computable function from the sentences of one to the sentences of the other, under which similarity of meaning is preserved. Then, we think there can't be such a theory unless L_1 and L_2 are compositional. The basic idea is pretty obvious. Suppose the languages are infinite and the translation theory is finite. Then presumably the translation theory works by first listing the translation relations among the primitive bases of the two languages (it says that 'chat' means the same as 'cat' and that 'chein' means the same as 'dog' and so forth) and then providing recursive procedures for constructing the translation relations among the infinitely many complex expressions. But, we suppose, this will only work if, in each of the two languages, the semantics of the complex expressions is determined in a regular way by the semantics of the primitive

[8] Reading 'part' as 'improper part'. So, in particular, the whole meaning of a syntactically primitive expression counts as 'part of' its meaning and as part of the meaning of any complex expression of which it is a constituent.

basis; i.e., only if the languages are semantically compositional.[9] If this line of thought is right, then the arguments that hold against similarity-based compositional theories (see above) are also arguments against similarity-based translation theories.

2.3 Intentional Explanations.

Paradigm intentional explanations appeal to principles that control the relations (rational, or causal, or both) between propositional attitudes with identical or overlapping contents. Thus, the 'practical syllogism' is supposed to be captured by some such schema as (S4).

S4. If a(gent) desires that P and believes that *not-P unless Q*, (where a believes that Q is contingent on an action that he believes himself able to perform), then, *ceteris paribus*, a tries to bring it about that Q.

Prima facie, (S4) is sound only if the formulas that substitute for P are identical or synonymous throughout. (It's plausible that weaker relations like, e.g., logical equivalence are insufficient even though they preserve the extensions or the relata.) Our point is that (S4) is patently *un*sound if substitutions for the schematic variables preserve only *similarity* of meaning.

Being a zerkon is similar to being a diamond; zerkons are manufactured to insure that this is so. But Marilyn Monroe did not think that zerkons are a girl's best friend, and she was right not to think so. Likewise, Chateau Rothschild isn't all *that* different from Chateau Plonk; but there are many who are prepared to pay for the one what they wouldn't pay for the other. Likewise, you may have a strong preference as to which twin you marry, even if the similarity between the twins is notable. *Etc*. None of these asymmetries of preference need indicate a palate that's excessively refined. To the contrary, the mediation of responses that discriminate sharply between similar situations is a fair part of what cognition is for. A landscape without a ravening tiger may be quite similar to a landscape with one. Tigers make a living out of that; it's part of the story about why they

[9] But what if translation turns out to be a *syntactic* relation? That is, what if the translation of L_1 expressions into L_2 expressions is fully determined given their respective syntactic structural descriptions (including, of course, their lexical inventories; see, e.g., Schiffer, S., Remnants of Meaning, Cambridge: Mass., MIT Press, 1987; Fodor, A Theory of Content, Cambridge, Mass., MIT Press, 1990, Chapter 7; Lepore, E., "Conditions on Understanding Language," Aristotelian Society Proceedings, 1996, pp. 41-60). This is perfectly ok with us, but notice that it is no help to somebody who says that translations *ipso facto* preserve similarity of meaning. The claim that syntax *determines* translation is not at all the same as the claim that syntax is what (good) translation *preserves;* what translation preserves, according to similarity based semantics, is presumably similarity of content. Well, if it's to do so in the case of translation between productive languages, infinitely many meaning similarity relations among complex expressions must all be projected from finitely many meaning similarity relations among primitive expressions. But, as we remarked in the text, it looks like meaning similarity is generally *not* preserved under the recursive processes that construct complex expressions from simpler ones; in general, the semantic property such processes can preserve is content *identity*. If this is right, a *syntactic* translation procedure could *not* be expected to preserve the *semantic* properties of expressions in the languages between which it operates if both languages are productive and the only semantic property defined for their expressions is meaning similarity.

have stripes. It may thus matter a lot which of two quite similar landscapes you are in. Your cognitive mechanisms are there in large part to help you to find out.

The underlying principle in all this is entirely obvious:

> If x and y are identical, then they are identical *tout court*. But if x and y are (merely) similar, then there is always a *respect* in which x and y are similar, and it can matter to their rational or causal relations what respect this is.

Here's another way to make much the same point. Presumably, $F = G \rightarrow \quad (F = G)$. But, of course, the corresponding conditional for similarity is invalid. The result is that the counterfactuals that identity statements guaranty are not, in general, consequences of the corresponding similarity statements; and we assume that 'P explains Q' is true only where 'if P, then Q' does support appropriate counterfactuals. (Likewise, *mutatis mutandis*, for 'P confirms Q'.) All this being so, it is hardly surprising that, in general, you can't replace content identity with content similarity in a semantic theory *salve* explanatory power.

There may be some way for a similarity based semantics to save the kind of intentional explanations that the practical syllogism gives in a content based semantics; but we don't ourselves know how; and we haven't seen any suggestions in the literature.

We end this section by remarking that it's our guess that nothing we've said so far would move Churchland in the least. What he'd reply, we continue to guess, is that we've failed to realize how very radical his revisionist proposals actually are: They are intended to abstract not only from the traditional idea that the central semantic construct is content identity, but also from the traditional inventory of problems that semantic theories are supposed to solve. Well, so be it. But we, in turn, aren't moved to exchange the old way of doing semantics for a new way on the grounds that the latter, though it will not solve what used to be thought the main problems about meaning, does promise to solve some new problems that it is unfortunately not now possible to enumerate. A pig in a poke is what it sounds like to us.

PART 2 – Why Churchland has no notion of content similarity.

Thus far our topic has been what role a similarity based notion of meaning might (or might not) be able to play in a semantic theory. Suppose, for the sake of the argument, that our doubts about this are misplaced and that an account of meaning similarity would be A Very Good Thing to Have. We turn to our second topic, which is why we don't think Paul Churchland's got one.

This requires a bit of background. Let's start with the notion of a semantic vector space. The basic idea is illustrated by Fig. 2.5 (after Churchland, P.M., The Engine of Reason, the Seat of the Soul, Cambridge, Mass., MIT Press, 1995, p.28),

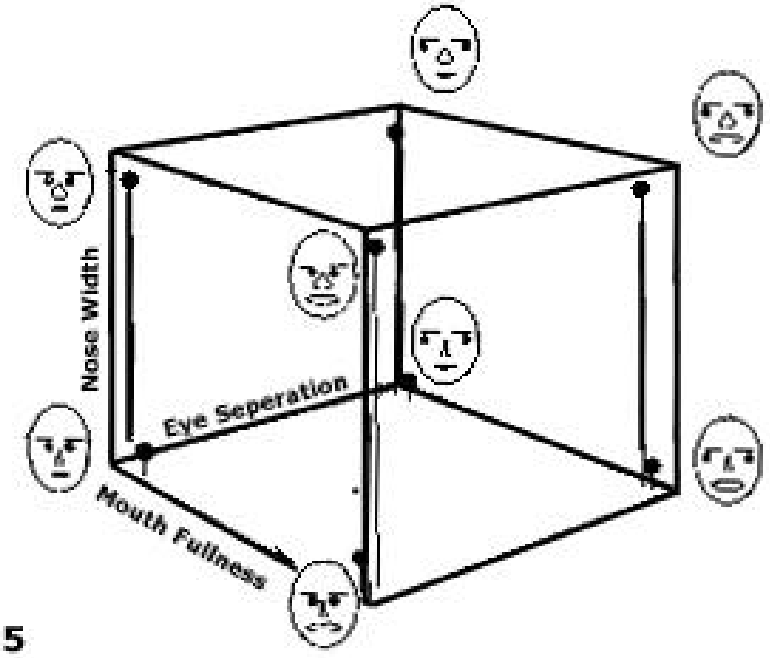

Figure 2.5

A rudimentary face space of three dimensions.

which is supposed provide a very simplified sketch of a representational system that might reconstruct the perceptual similarity of faces. Roughly, there is an n-dimensional space in which each representation of a face occupies a proprietary region. The dimensions correspond to properties of the faces that are salient for the perception of similarity (and/or for similarity judgements.) In the example, the space is assumed to be three dimensional for expository simplicity, "but this is highly unrealistic –our coding space for faces probably has at least twenty dimensions—but it does evoke the wide range of faces one can discriminately code with only a few resources"(Churchland, *ibid.*, p.29).) The intended interpretation of the notation is, in effect, that locating a given face F at a given position P on a given dimension D corresponds exactly to representing F as having the property D to the extent P. The empirical claim is that, given the right choice of dimensions, the propinquity of the items in a space predicts their perceived or judged similarity.[10]

Several quick points about this kind of representation:

–It doesn't matter that, in the example, a region of the space is identified with a representation (/concept) of a particular face. Any representation, either of an individual

[10] Churchland apparently thinks that the idea of developing a geometrical system of representation, in which concepts, contents and the like are identified with regions in a multidimensional semantic space, is novel and indigenous to connectionist models of mind. "[it] contrasts sharply with the kinds of representational and processing strategies that… cognitive psychologists…have traditionally ascribed to us" (Churchland, P. A Neurocomputational Perspective, Cambridge, Mass., MIT Press, 1991, p.171). In fact, semantic spaces have been around for quite a long while (see, e.g., Osgood, C.E., Suci, G.J. and Tannenbaum, P.H., 1967, The Measurement of Meaning, Urbana: University of Illinois Press). Most cognitive psychologists who worked with them eventually concluded that there isn't any question to which they are the answer. We think they were right to conclude that.

or of a property, would do equally well, so long as there is a natural and 'psychologically real' choice of salient dimensions along which to estimate the relative similarity of the things it represents.

–In fact, to claim that a concept can be properly represented as a region in a vector space really amounts to *no more* than claiming that there does exist a natural and motivated way of choosing such dimensions. Looked at this way, the differences between vector space encoding and more familiar accounts of meaning is easy to exaggerate. For example, a vector space encoding is entirely compatible with the traditional doctrine that concepts are 'bundles' of semantic features. Indeed, the latter is a special case of the former, the difference being that, whereas semantic features are generally thought of as having binary values, semantic dimensions are allowed to be continuous. Unsurprisingly, in light of this, semantic feature theories yield putative metrics of content similarity just as vector space theories do: *viz*., the relative similarity of two concepts corresponds to a (possibly weighted) sum of the features they have in common.

–The plausibility of identifying propinquities in a space with similarities of content depends entirely on the choice of the dimensions. If, for example, you change one of the dimensions in Fig 2.5 above so that what it expresses is not (e.g.) the width of a face's nose but the weight of the face's owner, the propinquities in *that* space will not predict judgements of perceptual similarity of the faces. This is the avatar, in the geometrical notation, of the perfectly general consideration that estimates of similarity must always be relativized to 'respects'.

So much, then, for the system Churchland wants to employ for annotating content relations among representations. Our main claim, in earlier phases of the exchange with him (see references, fn.2), was that moving to this notation requires rephrasing a number of traditional cruxes that arise for identity based theories of meaning, but that none of them are either resolved or bypassed. We briefly recapitulate the objections we raised.

As we've just seen, similarity of content is specified only with respect to a choice of the interpretation of the dimensions of a semantic space. There are three questions to be raised about such choices, all of which are patently counterparts to classical worries about individuating content in identity-based meaning theories.

i. How is the choice of semantic dimensions to be constrained? For example, what determines that we are to taxonomize faces by the width of the noses they contain, rather than (e.g.) noses by some relation among the faces that contain them? Or, more plausibly, why is it right (if it is) to taxonomize the apple prototype by reference to its location in a space of colors (prototypical apples are red) rather than the red prototype by reference to its location in a space of fruits (paradigm red is apple-red.) This is exactly parallel to the problem that arises in identity-based meaning theories about what constrains the choice of primitive features, as in: Why is it right (if it is) to define 'mother of' in terms of 'parent of' instead of the other way around?

Notice that this question is just as urgent for vector theories as it is for feature theories. For neither offers an interpretation of similarity relations, or indeed, of *any* semantic relations *among* the primitive features/dimensions. Rather, they purport to capture relations of similarity of concepts *relative to* the primitive features/dimensions (i.e., similarity in respect of the properties that the primitive features/dimensions express.) Hence, *ceteris paribus*, the bigger the inventory of primitive features/dimensions they posit the more semantic facts they fail to represent. In the limiting case, *every* concept corresponds to a primitive feature (i.e., to a dimension); and the theory becomes, in effect, disquotational. This sort of point hasn't, of course, gone unremarked in discussions of theories of meaning. The Old-Speak way to put it was: The notion of a semantic feature analysis is trivialized unless the notion of a semantic feature is constrained. Nothing about this changes when you go over from binary features to continuous ones.

We suggested, earlier in the exchange with Churchland, (Fodor and Lepore, Holism, *op. cit.*, 1992) that the reason this sort of problem seems not to bother him is that, like so many cognitive scientists, he believes deep down that only sensory properties can be semantic dimensions. This does solve the problem but at the cost of a recidivist and deeply implausible Empiricism. If Churchland has now got a better way out, we can't find it in his text.

ii. *What are the truth makers for claims about the semantic interpretation of the dimensions*?

The problem we identified for Churchland was that the only metaphysics of meaning he seemed to have on offer was one according to which the inferential roles of representations are the truth makers for claims about their semantic properties; and that could not be the right story about the semantic properties of the dimensions themselves.

As we saw above, to locate a concept (or a prototype) C with respect to a dimension D of a space is, in effect, to claim that the inference from being C to being D is constitutive (statistically or definitionally) of C. But if you thus identify the content of C by reference to its inferential relation to D, you cannot also identify the content of D by reference to its inferential relation to C. (You could, to be sure, identify the content of D with its position in some *higher order* semantic space; but that would initiate a regress.) This is, once again, exactly parallel to a classic worry about identity-based meaning theories: if the contents of complex concepts are identified by their inferential relations to primitive concepts, the content of the primitive concepts has to be identified in *some other way;* and nobody knows what other way that is.

As far as we can tell, Churchland is now clear that the notion that contents are somehow individuated by their inferential roles won't work for the content of the dimensions. Rather, he thinks, the semantics of the dimensions is somehow to be reconstructed in terms of causal relations between, on the one hand, the representations that have values on the dimensions and, on the other, facts about the world. We have no objection to his taking this way out; but he does now need some reason for not treating such causal

relations as the truth makers for *all* claims about the semantic properties of representations, thereby giving up, *inter alia*, the identification of contents with positions in semantic spaces. If (e.g.) the *redness* dimension has the content it does in virtue of causal relations between red things and the mental representation RED, why shouldn't the mental representation DOG have the content it does in virtue of its causal relation to dogs? Note the exact correspondence to a traditional question about identity-based meaning theories of the 'two factor' variety: If it is to be assumed that a one factor, causal theory works for the semantics of the primitive expressions, why is it to be denied that a one factor causal theory will work for the semantics of *all* the expressions?

iii. Strictly speaking, the vector theory identifies similarity of *content only with respect to concepts that have rankings on the same dimensions.* But it's implausible that this is true of the concepts people actually entertain. This raises a problem for similarity-based meaning theories that is exactly parallel to the traditional problem identity-based theories have about *how to avoid relativizing the individuation of concepts to the belief systems in which the concepts are embedded.*[11]

Suppose, for example, your concept NEBRASKA is just like mine, except in the following respect: You think that Nebraska is about a thousand miles from Guadalajara; whereas I, never having heard of Guadalajara, have no views about the distance between Nebraska and it. So, the situation is that your semantic space, unlike mine, has a dimension *proximity to Guadalajara* along which the concept NEBRASKA has a location. Question: How much (if any) of this kind of thing can be tolerated, consonant with our concepts of NEBRASKA being similar? We emphasize that the vector notation as such offers no answer to this question. Transdimensional *similarity* (similarity among concepts in spaces which differ in dimensionality) is undefined; just exactly as (lacking an analytic/synthetic distinction) an inferential role semantics offers no interpretation of concept *identity* for minds that don't share all of their beliefs.

Likewise, some dimensions presumably count more than others do in similarity evaluations, and the relative weight of the dimensions differs for different concepts. So, at least in principle, it's possible for minds to differ either in respect to the dimensions or in respect to their weightings, or both. Though you and I agree that literacy is a salient dimension for evaluating philosophers and engineers, I think it's important for philosophers to be well read, but it doesn't matter much whether engineers are; whereas you think it goes the other way around. Is there a principled answer to the question how (dis)similar this makes our concepts of engineers and philosophers? If so, what's the relevant principle?

There is, in short, a plethora of problems about how the similarity of concepts is to be evaluated across semantic spaces whose dimensionalities differ in various ways. These

[11] Maybe you don't mind if concepts are relativized to the belief systems in which they are embedded? So be it. But, of course, the semantic relativization problem lives just across the street from the semantic holism problem: If you have to relativize the identification of concepts to minds that have *some* of their beliefs in common, how (short of subscribing to an a/s distinction) do you avoid having to relativize the identification of concepts to minds that have *all* of their beliefs in common? Our point is that, if this worry is real, it persists unaltered when you change from a traditional semantics to a geometrical semantics.

are just the old problems about the relativization of concepts to beliefs; except that, in the new notation, talk about the way that a concept is positioned with respect to the dimensions of a semantic space replaces talk of the way that a concept is embedded in a theory. Whichever notation you prefer, you have the same problems about how to identify concepts and the like in the sorts of circumstances that typically hold *across minds* (or 'languages'; or 'theories').

We suspect that such problems are unsolvable unless the a/s problem isn't. Nonetheless, as far as we can tell, this is the issue to which Churchland takes his new theory to be germane. We're going to argue that this is a massive *ignoratio elenchi* on Churchland's part. The proposal he has on offer isn't *of the right logical type* to be a solution of the problem of transdimensional content individuation; and, though there is a problem in the general vicinity that it *is* of the right logical type to solve, it doesn't, in fact, solve that problem either.

Churchland's proposal co-opts work by Laasko and Cotrell, ("Qualia and cluster analysis; assessing representational similarity between neural systems", unpublished ms, 1998) which, he says, "constitute[s] a decisive answer to Fodor and Lepore's challenge…. [Laasko and Cotrell] successfully deploy one member of a large family of mathematical measures of conceptual similarity, measures that see past differences –even extensive differences..." (Churchland, P. On the Contrary, Cambridge, Mass., MIT Press, 1998, p.81) and now, given the nature of the challenge that we had offered (*viz.,* that there's a relativization problem for vector space semantics just like the relativization problem for identity based semantics) one expects this sentence to end "of the *semantic* dimensions of the conceptual spaces". For, to repeat, a vector space semantics identifies concepts by reference to their positions in a semantic space; and the problem we raised—the one that Churchland apparently thinks Laasko and Cotrell have solved—was precisely 'how can minds that don't have the same semantic spaces have the same (or similar) concepts?' Here, however, is the way the sentence that we just quoted actually does end: "measures that see past differences… in the connectivity, the sensory inputs and the *neural* [our emphasis] dimensionality of the networks being compared" (Churchland, *ibid*., p.81).

Now, something has gone wrong here. Neural spaces and semantics spaces are quite different sorts of things. As we've been seeing, positions along the dimensions of the semantic space relative to which a concept is situated correspond to propositional attitudes the holding of which is supposed to be constitutive of having the concept. Whereas, the dimensions in a neural space *correspond to neurons;* and the positions along a given dimension specify the degree of excitation of the corresponding neuron (at a time).[12] Unsurprisingly, though he isn't reliable in observing it, this difference does show up in Churchland's notation now and t hen; see Fig.4.22 (from Churchland, The Engine of Reason, *op. cit.,* p. 88) notice that the dimensions of *this* cube are labeled with the names of neurons, not with the names of semantic dimensions (compare Fig. 2.5 above.)

[12] If we simplify (as, indeed, connectionist models often do) by assuming binary dimensions as opposed to ones with real number values, a state description relative to a neural space is an assignment of 1 or 0 to each neuron depending on whether it is active (at t).

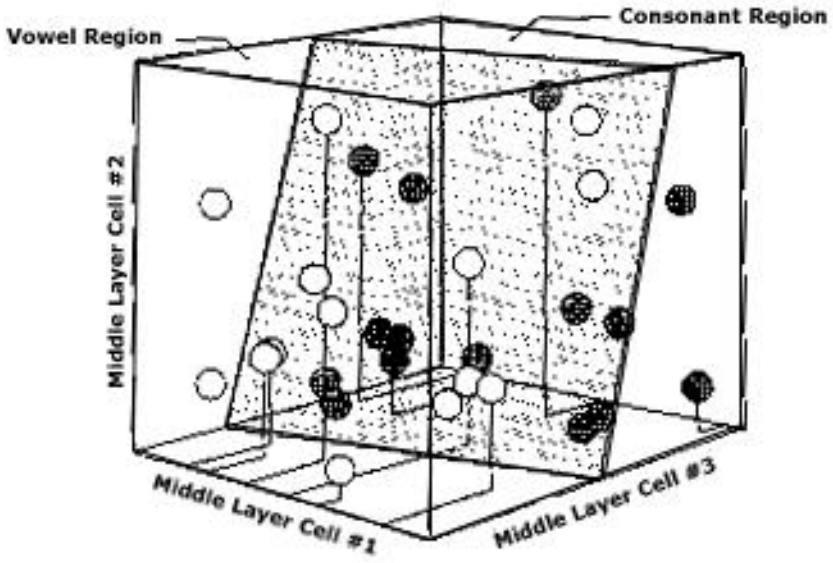

Figure 4.22

The activation space of the cells at the middle layer of NETtalk positioned in the space corresponds to the seventy-nine distinct letter-to-phoneme transform actions.

Now, just as there is a relativization problem about the identity (or similarity) of *concepts* and the like across *semantic* spaces of different dimensionality, so too there is a relativization problem about the identity (or similarity) of *brain states* across *neural* spaces of different dimensionality. For example, if you are a connectionist, you identify circuits with n-tuples of nodes and links; and the nodes and links in a network are themselves identified *by the totality of their relations to one another*. On this picture, the notion 'same (/similar) circuit' is defined *only* for networks of identical dimensionality, just like the notion 'same (/similar) content.'

We think that this analogy between individuation problems has caused Churchland to confuse providing a transdimensional notion of brain state identity with providing a transdimensional notion of content identity; it's because he has failed to observe this distinction them that he seems indifferent whether the dimensions of the spaces he talks about are semantic or neural. It's likewise because of this confusion, and because Churchland thinks (wrongly; see below) that Laasko and Cotrell have solved the transdimensional individuation problem for brain states, that Churchland believes (again wrongly) that Lassko and Cotrell have solved the transdimensional individuation problem for content states. So we will now argue.

Let's do this slowly. There is, to repeat, a problem about type individuating brain states across neural spaces of different dimensionality,[13] just as there is a problem about

[13] For that matter, there is a problem about typing brain states across neural spaces with the *same* dimensionality if the dimensions of the spaces are supposed to correspond to neurons as *anatomically* individuated. There really is, we think, a dilemma hereabouts. If a neuron is a *physiological* unit (e.g., neurons are individuated by their connectivity and the like), then there is a holism problem for them just as

individuating contents across semantic spaces of different dimensionality. But *it can't simply be taken for granted that a proposed solution for the first of these problems would solve* (or even help to solve) *the second.* Suppose you have some scheme for picking out homologous regions across neural spaces of different dimensionality and that you're convinced that it preserves similarity (or identity) of neural types. That is, homology is defined so that if there is a region of an n-dimensional space such that:

(1) all the brain states in this region are neurologically similar or identical; and if
(2) for every space of m-dimensions ($n \subset m$), if a brain state has a location in the n-space it has a location in the homologous region of the m-space; then:
(3) the brain states in the homologous regions of the m-space are likewise neurologically similar or identical.

That is, of course, a very strong condition to place on a notion of transdimensional homology for neurological spaces. But our point is that, even if you had a notion of transdimensional homology which, in this sense, preserved *neural* state similarity (or identity), you'd have no right to assume that it likewise preserves *semantic* similarity (or identity). If you want to claim that it does, you will need to provide an argument. Roughly, you need to show why, if the brain states located in a region of a certain neurological space are similar or identical *in content*, then so too are the brain states located in a region of any higher-dimension neurological space that is homologous under the preferred trandimensional mapping.[14]

Notice that taking physicalism for granted (which, for the sake of the argument, we're happy to do) does *not* provide the argument that's required. 'Token' physicalism says that if individual brain states are identical, then so too are the corresponding individual

---

there is for blatantly functional entities like circuits. On the other hand, if neurons are *anatomical* units, there is no guaranty that brains that have the same neural populations are ever in the same states by any functional criterion. As far as we can tell (e.g., from the labeling of his Fig. 4.22) Churchland prefers the first horn of this dilemma; i.e., he (implicitly) takes the individuation of neurons to be sensitive to their functions. This raises, but does not solve, the question how exactly the dimension labels of his neural activation spaces are to be interpreted; i.e., *which* functional properties 'being neuron A' and the like express.

Begging the question whether, or to what extent, the 'neurons' in their networks are supposed to be *anatomical* as well as *functional* units is a standard practice in connectionist texts. It makes it seem to uninitiates that connectionism is brain science, not just computational psychology. That is very good for getting grants, but it does cause a lot of confusion.

[14] It is worth noting that the claim that content state types correspond to brain state types under some or other criterion of individuation of the later is implausibly strong quite aside from issues of *transdimensional* individuation. In particular, such a claim would require that wherever states of mind overlap in their intentional contents, there is also an overlap of the corresponding brain states. So, for example, there would have to be a (constituent) brain state common not just to tokens of the thought *that's red*, but also to tokens of the thought *I have decided to paint the cat red*; and to tokens of the thought *red is as far from green as blue is from yellow*; and so forth. There is, as far as anybody knows, not the slightest reason for supposing that the relation between mental contents and brain states meets this constraint even if issues of transdimensionality are ignored. In fact, as things now stand, there are *no* known examples of cases where it does. It's a nice irony that it fails notoriously to be satisfied by the relation between circuits in connectionist networks and the contents that the circuits are supposed to encode. That's exactly why connectionists have so much trouble with compositionality.

intentional states; but token physicalism says nothing about whether there are such things as transdimensional counterparts of brain state tokens; or whether, if there are, they *ipso facto* share any of their semantic properties; or, if they do, which ones. Likewise, 'type' physicalism guarantees that if brain states are of the very same kind, then so too are the corresponding content states. But it makes no commitments about what, if any, content relations obtain for brain states that are merely *transdimensionally homologous* according to some or other typological scheme. Churchland must be assuming a typology of brain states that meets the condition that when neurological similarity (or identity) is preserved under a transdimensional mapping, content identity is preserved too. Nothing else would warrant his persistent practice of putting *semantic* labels on the dimensions of *neural* spaces and *vice versa*. But, to repeat, that semantic similarity is preserved transdimensionally when neural similarity is preserved transdimensionally doesn't follow from any kind of physicalism we've heard of, nor would it seem to be remotely plausible on independent grounds.

Here, to summarize, is the dialectical situation as we understand it so far: Churchland thinks that Laasko and Cotrell have provided a notion of 'kind of brain state' which preserves similarity of neural state types across spaces which differ in their neural dimensionality. And he (tacitly) infers that they have *thereby* provided a transdimensional notion of content state similarity as well. But, to put it mildly, this inference is tendentious. Churchland needs an argument for drawing it, but provides none. We doubt that he has one that he's shy about producing. Rather, our guess is that he has simply confused *semantic* spaces, (of which the typical occupants are intentional states, so described,) with *neural* spaces, (of which the typical occupants are brain states, so described).

OK so far. We now propose to show that, in fact, the sort of criterion Laasko and Cotrell suggest for type individuating brain states across dimensionality differences pretty clearly does *not* preserve either identity or similarity of the contents of the states.[15] Basically, Churchland's idea is that, even if my neural space has fewer dimensions than yours, we can embed the one in the other so long as the dimensions of my space are a subset of the dimensions of yours. In which case, we can assign each of my (actual or possible) neural states to the same state type as some (actual or possible) neural state of yours. *A fortiori*, we can type identify (at least some) brain states across (at least some) differences among neurological spaces. That, as far as we can tell, is the burden of the following remark:

> If a pair of dimensionally diverse spaces both contain … a same-shape n-dimensional hypersolid…, then both spaces must each either constitute or contain an n-D subspace, an n-D hyperplane of some orientation or other, that confines each hypersolid at issue. Since those respective hyperplanes… have the same dimensionality, there will be no problem in comparing the two solids. (Churchland, On the contrary, op cit, p.91).

[15] *A fortiori*, if the Laasko and Cotrell sort of view is right about how to individuate brain state types, then type physicalism is false about the relation between brain state types and intentional state types. (Type physicalists should not be alarmed by this however; we'll see presently that their proposal for transdimensionally individuating brain state types doesn't work either.)

We paraphrase (under correction, to be sure):

> *Claim Q:* If my neural space is properly embedded in yours, then (*ceteris paribus*) my (actual or possible) brain states are a subset of yours.

Suppose, for the moment, that Q is true; hence that if a neural space of lower dimensionality is conservatively embedded in a neural space of a higher dimensionality, then each brain state that has a location in the former space has a corresponding location in the latter. Our argument will be that even if Q is true *it doesn't provide for the transdimensional individuation of content states.* That's because *R* –which is the analog of Q for *semantic* spaces—is clearly false*;* content identity is *not* preserved under the embedding of one semantic space in another.

> *Claim R:* If my semantic space is conservatively embedded in yours, then (*ceteris paribus*) my concepts are conservatively embedded in yours.

Here's a candidate counterexample to R (the Nebraska/Guadalajara case discussed above is another.) Suppose I have only three semantic dimensions Hard-Soft, Black-White, Heavy-Light. And suppose my concept ROCK is identified with a vector that specifies a region in the space that these dimensions define. Likewise for you: Your semantic space also has these dimensions (*inter alia*); and you too have a concept that lies within the boundaries of the space that they define. The difference between us is that, whereas these dimensions define *the whole* of my semantic space, they define only an embedded subspace of yours. Now, *does it follow that our concepts ROCK are similar?* We shouldn't have thought so. For, perhaps your space has a dimension ANIMACY which, by assumption, mine lacks. And suppose that you think that rocks are actually quite animate; much more animate, say, than turtles; at least as animate, indeed, as half the members of the D.A.R. Is it *still* the case that your ROCK concept is similar to mine? If so, suppose also that your space contains a dimension for abstractness, and that you think that rocks are pretty abstract; at least as abstract as the natural numbers, say, though maybe less abstract than the real numbers. Is your concept ROCK *still* similar to mine? *If there are any such cases where the right answer is 'no', Churchland looses.*

Our point is that, even if concepts, semantic properties, intentional states and the like, can be identified with positions in a state space, and even if position in that state space is preserved when the latter is embedded in a space of higher dimension, it doesn't follow that those (or indeed any) semantic properties, intentional states, etc. can be identified with positions in the higher order space. That's because *the conditions for the individuation of contents are non-monotonic.* This too is something that everybody already knew; indeed, it's just the a/s problem in one of its familiar guises. Presumably Homer has a concept of water which was similar to ours *as far as it went.*[16] (Water is wet, it's potable, it freezes in the winter…and so forth). There remains the question whether Homer's concept of water was similar to ours *tout court,* and the relevant considerations are much the same one's that the corresponding question about the *identity* of our

[16] Assuming, as Churchland of course does, that what concepts you have depends on what beliefs you have.

concepts would turn on: Our WATER concept is embedded in a theory (a *Weltanschaung*; a form of life; blah, blah) many of whose dimensions are quite different from any of the ones in Homer's semantic space. What, if anything, follows about which concepts we share with him?

Recapitulation: even if we did have a transdimensional notion of 'brain state similarity' it wouldn't follow that we'd then have a transdimensional notion of 'content similarity.' In fact, there is no reason to suppose that the neural similarity of brain state types corresponds to similarity of their content state types under *any* mapping scheme for which there is empirical warrant.

We close by remarking that Churchland's sort of story really doesn't work even for the transdimensional typing of brain states. Q is false because the identity conditions for neural state types, like the identity conditions for content state types, is non-monotonic.[17] A brain state that is identified with a position in a neural space of *n* dimensions need not be type identical with *any* brain state that is identifiable in an *m*-dimension neural space in which the *n*-space is embedded. This is because brain state types are *functionally* individuated; and functional individuation is itself non-monotonic.

Suppose we have a neural space of *n* dimensions; and suppose that the function of a certain activation state of these neurons is to cause a flight response. Now embed the *n*-space in a one of higher neural dimensionality. (Since each dimension of a neural space corresponds to exactly one neuron, this requires putting the original n-neurons in a brain that is m-n neurons bigger, preserving the connectivity of the former to one another). It surely doesn't follow that there is any activation state of these (or, indeed, of any) neurons in the larger brain whose function is to cause flight responses. Indeed, it doesn't even follow that activation states defined over the original *n*-neurons have *any function at all* in the larger brain; that depends, *inter alia*, on what *other* neurons they are connected to. It is, for example, perfectly possible there is *no* physiologically distinguishable system that all or only the original *n* neurons belong to when they are in the bigger brain.[18]

Brief conclusion: The question we were worried about in our earlier exchanges with Churchland was: 'How can there be *content similarities* between minds that have different beliefs?' We were worried about this because Churchland keeps saying that similarity semantics avoids the main problems that a/s semantics is plagued with, and our understanding is that 'How can there be content identities between minds that have different beliefs?' *is* the main problem that a/s semantics is plagued with.

---

[17] We are not, please note, asserting – or even suggesting—that content individuation is non-monotonic because it's a species of functional individuation. Heaven forfend.

[18] Suppose, however, that it's denied that the individuation of kinds of brain states is functional. Even so, it's far from obvious that either identity or similarity of brain state kinds would be preserved under the sort of transdimensional mapping Churchland imagines. It's patently not the case that brain states that are similar *in a given world* because they share certain nonfunctional properties are *ipso facto* similar in *any* world in which share those properties. There *may* be ways of choosing canonical neural dimensions that meet this condition; but there's no a priori reason to believe that there are.

But the question Churchland actually offers an answer to is, 'How can creatures be in similar brain states if they have different numbers of neurons?' This is an interesting kind of question, to be sure; but it has nothing in particular to do with meaning. To the contrary, it arises for any theory that is committed to the functional individuation of things that it quantifies over. This is so whether or not the things that it quantifies over are psychological/intentional/semantic. In particular, the question of transdimensional functional state individuation arises for any theory that is committed to kinds of states that can have multiple realizations. Cf., 'How can tubes and transistors both be amplifiers?' that 'How can jets and diesels both be engines?' 'How can a wood thing and a plastic thing both be pawns?' and so on.)

It would be nice to have a transdimensional, similarity preserving, type-identity criterion for brain states, and it would also be nice to have a transdimensional, similarity preserving, type-identity criterion for content states. And it would be *VERY* nice (and *VERY* surprising) if the two criteria turned out to be the same. But, as far as anybody knows or has any reason to believe, these two individuation problems raise quite different ontological issues, and their solutions – if, indeed, they have got solutions—are likely to be largely independent. All they have in common, as far as we can tell, is that both kinds of individuation are non-monotonic, and that Churchland hasn't got a workable account of either; and nor, of course, has anybody else.