On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought[*]

(appears in *Synthese*, Vol. 117, No. 1, 1998)

Robert D. Rupert

I. Introduction

In an attempt to explain the nature of mental states and their role in the lives of human beings, philosophers often begin with assumptions of the following sort: (1) there exists in the human mind/brain a system of representations (sometimes called 'terms in a language of thought', or 'terms in LOT')[1] among which there are causal connections; and (2) the causal connections holding between mental representations partly explain human behavior. In order that assumptions (1) and (2) be wholly legitimate, at least the two following questions should be answered: How are mental representations individuated? and How do these representations acquire intentional content,[2] the possession of which qualifies them as representations and is supposed to give substance to the claim that the presence of representations helps explain human thought and behavior? In this paper, I aim to show that what appears to be the most promising answer to the first question, about individuation, has important implications regarding what might count as an acceptable answer to the second, concerning the determinants of mental content.

While assumptions (1) and (2) are often made by philosophers of other theoretical orientations, the arguments in this paper specifically address attempts by naturalistically inclined philosophers to understand mental representation. Naturalism in the philosophy of mind holds that the human mind is a completely organic, biological phenomenon. Furthermore, what makes naturalism a distinctive philosophical orientation is its commitment to the methods employed by the natural sciences in investigating any and all parts of reality, including, of course, the human mind. So naturalism implies not only a methodology, more on which below; it also implies that

our theories of individuation and content for LOT terms should be spelled out in nonsemantic terms.[3]

In the remainder of the paper, I proceed as follows. After some brief remarks on naturalistic methodology, I outline a view of how to nonsemantically individuate mental representations, a view currently emerging from research that treats the mind as a dynamical system. Here I note the promise of the dynamical systems-based view to provide a better picture than we've had to date of how a cognitive system, one that employs representations, can be integrated with the other component systems (e.g., the perceptual systems or the muscular system) of the human subject. While I present some reasons for thinking that we should adopt the dynamical systems-based approach to the study of cognition, it is not my intention to provide detailed and weighty arguments in support of this approach. My goal is to present enough of the dynamics-based view to understand what its implications might be for a theory of content for terms in LOT. In particular, I wish to bring the dynamics-based view of mental representation to bear on the debate over the viability of a naturalistic semantics for LOT terms. I argue that given the dynamical systems-based view of mental representations, and given a certain view of the naturalistic philosophical project, the defender of a naturalistic semantics for LOT terms should question the relevance of many thought experiments often believed to be respectable philosophical coin in discussions of mental content. The striking result here is that a subject's actual causal history may play a larger role in determining the content of her LOT terms than has been widely thought.

While I present the argument in more detail in later sections, it may be helpful at this point to preview my basic line of reasoning. Typically, a thought experiment requires our imagining a counterfactual situation, and normally, it is metaphysical possibility alone that constrains the framing of the counterfactuals that describe this hypothetical circumstance. In contrast, I argue that the dynamics-based view of the nature and the acquisition of mental representations, together with a thoroughgoing naturalism, places more stringent requirements on the acceptable framing conditions for those counterfactuals that are the very stuff of philosophical thought

experiments about meaning. If a dynamical systems-based theory of individuation conditions for mental representations tells us that a given LOT term *t* cannot, physically speaking, be tokened under circumstances *c*, then it is fruitless for a naturalist to ask, as part of a philosophical thought experiment, what our intuition-based attribution of content would be were *t* to be tokened in *c*. Accordingly, we should reevaluate the role of thought experiments in critical discussions of naturalistic semantic theories for LOT terms. Upon reevaluation, we find an increased plausibility attached to theories of content according to which an LOT term's reference is fixed by the subject's actual history, this because it is much more difficult to give legitimate counterexamples to actual history-based theories.[4]

II. Remarks on Methodology

In basic agreement with Michael Devitt (1994, pp. 565-9),[5] I assume that when faced with (what we suspect is) a natural phenomenon, our investigation of that phenomenon should begin with an examination of its most widely recognized or most obvious instances. Thus, if we wish to understand representation as a natural phenomenon, we should begin by inspecting the best examples we have of beings who use full-fledged representations: we should examine the beings in whose use of representations we are most confident, i.e., human beings.[6] As Devitt rightly points out, even our strongest intuitions as to which cases best exemplify being an F (in this case, F = a system that employs representations) are defeasible, in principle (Devitt 1994, p. 568). But as Devitt also notes, our strongest intuitions regarding which items have property F should serve as the starting point of our attempt to figure out what it is to have F. The upshot, then, is that the naturalist should endeavor to understand representation as it actually exists in humans before making any general, species-independent claims about representation or intentionality. (Of course, results in ethology and animal physiology can inform our study of human mental representations; my point is that the latter field takes precedence. While this position may strike some as unnecessarily chauvinistic, it seems more plausible when we recognize that the very human abilities [e.g., flexible use of language, creative recombination of concepts, thoughts

about the past or future] that led early cognitive scientists to reject behaviorism and take internal mental representations seriously are abilities that seem to be lacking for the most part in nonhuman species.)

We cannot avoid the question of individuation when developing a naturalistic theory of mental representation: if we wish to say that a term in LOT acquires semantic value in virtue of certain of its nonsemantic relations, we should have in mind some characterization of what one of those terms is and what sets it apart from the others (i.e., some characterization other than 'the term with such and such content'). Before moving on to examine the relation between dynamical models of the mind and LOT term individuation, note the limitations of Devitt's suggested methodology were we to attempt to apply it to the question of individuation. As Devitt describes it, naturalistic semantic theory should take as initial data points our semantic intuitions as to which items have specific semantic properties (A semantic property is, for example, meaning that *p*, where *p*'s value is to be some specific content). However, to construct a complete naturalistic semantic theory for LOT terms, we need a detailed theory of the nature of the items that are supposed to bear these semantic properties. Here humans would seem to have nothing analogous to their semantic intuitions, i.e., nothing that informs them pretheoretically of the nonsemantic individuation criteria for LOT terms. When a human thinks of water today, how is she to know that she is tokening the same LOT term she tokened when she thought about water yesterday? Contrast this with the case of term individuation in a natural language. Knowing nothing about the meaning of two, token, spoken words, a person can make a fairly reliable guess as to whether the two are instances of the same type, especially if these two words are from a language she speaks. With respect to terms in LOT, though, there seems to be nothing analogous to the capacity for auditory discrimination that underlies judgements about the similarity of spoken words whose meanings are not known. Thus, in developing a nonsemantic theory of individuation for LOT terms, we cannot take thinkers' intuitions as the initial data points for our theorizing (the way, for example, a linguist might in identifying the range of acoustical variation speakers of a given language are willing to allow in the production of a

single phoneme).[7] We should look elsewhere, then, for individuation criteria for LOT terms, and dynamical systems theory appears to have much to offer in this regard.

III. The Dynamical Systems-Based Approach

We would like to be able to explain how the various LOT terms differ from each other without making reference to their content. We might reasonably begin by cataloguing the relevant free parameters of the human physical system in which the LOT terms appear. By 'free parameters' I mean the independent, noncognitive dimensions along which the state of the relevant portion of the system can vary; of these parameters, we should be particularly interested in those whose variations we suspect make a difference to the individuation of LOT terms. If we were to take the idea of a *language* of thought too literally, we might identify the relevant free parameters as the letters out of which LOT terms are constructed; 'cat' in LOT would then be different from 'bat' in LOT in virtue of the terms' having distinct initial letters. This approach does not seem promising, however, given our general failure to find such letters written in the brain. More plausibly, we might consult contemporary neuroscience as a way to locate properties of the brain on the basis of which we can individuate a subject's LOT term types. Here the set of relevant parameters would seem to consist of the firing rates of neurons, together with values of other parameters, such as the brain's temperature, that affect the way in which the nervous system's components operate.

Note that it is incredibly difficult to construct a model of cognition that takes into account *all* of what we might suspect are relevant parameters. For this reason most cognitive theorists offer simplified constructs meant to capture something important about cognition at the risk of excluding some pertinent details. Consider the fact that there are approximately one trillion neurons in a human brain (Stillings et al. 1987, p. 267, Churchland and Sejnowski 1992, p. 51). If we begin our cognitive theorizing treating rates of neural firing as our free parameters, then idealization and simplification seem unavoidable.[8] This is true even of connectionist models, which are supposed to be anatomically realistic in a way that other models of cognition are not:

such networks typically consist of a small number of units relative to the number of neurons in the brain.

Having decided how to characterize the free parameters of (the relevant part of) a given system, we can employ the tools of dynamical systems theory to analyze that system. Begin with the idea of state space. The state space is an abstract collection of possible overall states of the system, with one state corresponding to each possible combination of values of the free parameters. Consider a standard light switch: to keep things simple, imagine that the switch can be in only two positions, up (on) or down (off); accordingly, the system's state space consists of the two possible states of the system, one where the switch is up, and one where it is down. In addition to a system's state space exists a mathematical space called a 'phase space' (Frequently this is also referred to as the 'state space', leaving context to disambiguate when necessary). The phase space is an *n*-dimensional mathematical space, where *n* equals the number of free parameters in the system, the various combinations of values of which each represents a possible state of the system; the phase space specifies, by means of a vector field, how the system will evolve from any given state (The pattern of distribution of vectors in the field is often referred to as the 'shape' of the phase space or as the 'phase portrait'). The phase space's shape is determined by a set of differential equations (or difference equations in systems that move through discrete states) that define what are called the 'dynamics' of the system. The light switch's phase space consists of one dimension that can assume only two values, the phase portrait of which will likely bore us to tears: as I've described it, the switch is essentially a static system with two discrete states; in the absence of disturbance from outside the system, the switch simply stays in the state in which it begins. The equation governing the light switch's behavior takes as input the initial state and yields as output that same state (though if it is to accurately capture the situation of a real light switch, the equation must also include an expression allowing for influence from the outside, which complicates matters); thus, as a closed system, its phase space is no more than two fixed numerical points.[9]

Van Gelder and Port (1995) lay out many reasons for thinking that the application of dynamical systems theory to the cognitive system will lead to important insights into the nature of human thought; in doing so, Van Gelder and Port argue for a dynamical systems-based (DS)[10] view of cognition.[11] They are at pains, and for good reason, to explain exactly what advantage is gained by cognitive theorists when they employ the tools of dynamical systems theory rather than those of the "old-fashioned computational approach". (That they give such reasons is important, for many philosophers are inclined to think of DS models as describing how the central nervous system instantiates a mind, without thinking that DS models have much to do with the nature of cognition.) Of Van Gelder and Port's concerns, most germane for present purposes is their worry that the computational theorist cannot bridge the gulf between physical stimuli at the sensory periphery and the abstract, symbolic entities that sensory transductions produce and over which computational processes are defined (Van Gelder and Port 1995, pp. 26-30). The activity of sensing the world around us begins as a dynamical interaction with that world. On the DS approach, there is no need to identify the point at which electrical impulses in the nervous system become abstract symbols, for the numerical values assigned to firing neurons at the sensory perimeter (or groups thereof) are values along dimensions of the same phase space in which one finds LOT terms. Thus, on the DS view, there is no "theoretical gap between cognitive systems and their surrounds" (Van Gelder and Port 1995, p. 28), as there is on the computationalist's view. Taking the DS approach, the mathematics of cognition are continuous with the mathematics of transduction. For this reason, DS theory yields a deep understanding of how LOT terms are introduced into the cognitive system by treating them as modeled after or continuous with the brute sensory signals that are often the causes of tokenings of LOT terms. Furthermore, it follows from the methodology endorsed in section II that if we are to discover the nature of mental representation, we should seek first to understand how LOT terms appear in the human physical system; for *physically embodied* people are the best, perhaps the only, examples we have of full-blooded representational systems.

To some readers, it may seem odd that I talk approvingly about the DS approach while continuing to speak of mental representations, this because many cognitive scientists who advocate treating the mind as a dynamical system have an eliminativist bent with regard to mental representation. I shall consider some eliminativist claims in more detail in the closing section, after first explaining what I take to be the relevance of the DS approach to naturalistic semantics. For the time being, however, I hope the two following points will suffice: first, it should be noted that not all cognitive scientists in the DS camp embrace eliminativism;[12] furthermore, there is good reason to be suspicious of the sweeping eliminativist claims that sometimes issue from those DS theorists who doubt the value of continued talk of mental representations.[13] Thus, I provisionally adopt the view that the DS approach can to tell us something important about mental representations, other than that there are none.

IV. DS Views, Causal History, and the Introduction of LOT Terms

To set the stage for the remainder of the discussion, I now indulge in a minor digression, which may prevent a variety of possible misunderstandings and consequent objections. Let us distinguish between what I will call 'semantically basic' LOT terms--terms that Fodor (1990a, p. 92) describes as having their content fixed in an atomistic fashion--and those LOT terms whose content is fixed partly by the *content-laden* contribution of other mental states. Two contrasting examples serve to illustrate this distinction: On one hand, it's plausible that the infant's LOT term 'object' emerges very early in development and acquires its reference independently of any content-laden contribution of other mental states.[14] On the other hand, recall Putnam's example of speakers' intentions to use the English term 'water' to refer to all stuff of the same liquid kind as the samples confronting them during the period when the term 'water' is being introduced into the language (Putnam 1975). Here the *content* of the term 'liquid' plays an essential role in limiting the representational content of 'water' to H_20. Assuming that a roughly analogous process sometimes fixes the content of terms in LOT, we will have cases where LOT terms acquire determinate extension in virtue of the contributions made by other representations that

already have their extensions fixed (In fact, this would seem to be the case for most LOT terms). It is beyond the scope of this piece to make precise this distinction between semantically basic mental representations and those whose contents depend upon previously established content (See Bartsch 1996, p. 424, for a more rigorous description of what would seem to be a similar distinction). However, two comments are in order: First, I do not mean to be suggesting that semantically basic mental representations serve as definitional simples out of which all other content is constructed. My point is only that the intentional content of the semantically basic representations is fixed independently of the content of other LOT terms, and that once their content is fixed, such content can contribute to the fixation of content of other LOT terms. Second, it is of paramount importance to bear in mind that the points I make in the remainder of this paper apply most clearly to semantically basic LOT terms, even when examples chosen for ease of exposition suggest otherwise (as with the discussion of 'horse' in section VI).

In his development of DS models of natural language processing, Jeffrey Elman identifies LOT terms with regions in a state space (Elman 1995). This way of putting things is a bit misleading, however.[15] Which LOT terms a subject can be said to *possess* depends on more than a simple partitioning of the state space into quadrants or some such; it depends also on the phase portrait. For a subject to possess an LOT term, a certain point (or region) in the subject's phase space must have the property of being an attractor in the phase space: it should be a point (or region) p in the phase space such that when the system passes through states that correspond to points near p (i.e., points in what is called p's 'attractor basin'), the system heads toward p. Generally speaking, we should limit our identification of a given subject's LOT terms to attractors because the property of being an attractor confers upon a point a special status in the dynamics of the cognitive system: There is no guaranty that the system will ever reach a state that corresponds to a given attractor; however, there is a much greater likelihood that the system will come close to, pass through, or remain fixed at a state corresponding to an attractor than there is that the system will come close to, pass through, or remain fixed at a state that

corresponds to a randomly chosen point in the phase space. According to the DS view, the cognitive system typically moves from one attractor to another (perhaps with periods of chaotic behavior in between). It would seem ill founded, then, to identify a specific set of coordinates in the phase space as corresponding to an LOT term simply because such coordinates are there to be identified. Even were we to associate one of these points with an LOT term, we would not want to say that the subject possesses that term or that the subject understands the concept associated with that LOT term. A dynamical model of speech comprehension can, for example, identify a region of a subject's state space as the 'electron' region. But if the subject has never heard of electrons or even entertained the idea of subatomic particles, then even if an experimenter says to the subject "think about electrons", the subject's cognitive system will not pass through what we might want to call the 'electron' region in the subject's state space. The subject might pass through a state in something like her phonological representation space that we can identify with her recognition of the English word 'electron'; however, this would not be the same as having acquired the LOT term 'electron' as it's normally conceived of, i.e., as the concept of an electron. Thinking in terms of the phase space again, there simply is no attractor that we would identify as corresponding to the LOT term 'electron'. This term is missing from the terrain of the phase space.[16]

Given our interest in actual history-based semantic theories for LOT terms, we should want to know how attractors appear on the scene, i.e., how the phase space that models the cognitive system of a particular subject takes on its shape. The research carried out by Esther Thelen and her associates (see the various studies cited and described in Thelen 1995) offers one DS view of the development of the phase portrait, a view which takes the development of motor skills in infants as fundamental to cognition. Thelen begins with the assumption that the child's development of motor skills can be identified with the emergence of certain attractors in the phase space that corresponds to the range of the child's physical movements. Thelen then shows how these attractors develop as responses to infants' concrete interactions with the world.[17] Natural patterns of movement are slowly altered allowing the infant to achieve her goal (e.g., the

grasping of a toy). Thelen further explains how these concrete interactions might serve as the basis for the development of more abstract concepts, such as that of acting in general (Thelen 1995, pp. 95-98). A related DS perspective comes from Jean Petitot. While the emphasis here is not on developmental issues, Petitot illustrates the way in which abstract structures (for example, concepts of semantic roles) can be derived from visual scenes (Petitot 1995). Petitot's work bears on current concerns because certain tools necessary for human cognition, i.e., the nonsemantically individuated LOT terms that can serve as representations of abstract properties, seem again to be emerging from the actual physical experiences of the subject.[18]

Assume that we acquire many of our most abstract and fundamental LOT terms, e.g., 'object', in the ways described by Thelen and Petitot. It seems likely, then, that the emergence of a given, semantically basic LOT term *t* is highly dependent on the subject's having had certain kinds of experiences, i.e., the subject's having had a certain kind of actual history. This is not to say that during the process of acquiring *t* there has to have been anything approaching a perfect covariance between *t* and the members of the extension that we think should be assigned to *t*. Yet, in the absence of perfect covariance, it may be that the only way for the subject to acquire *t* (the only way for the subject's phase space to take on a shape such that we are willing to say that the subject has acquired *t*) is for the person to have had reference-fixing contact with *t*. I do not here offer an account of this reference-fixing relation.[19] My focus on the subject's actual history is only meant to show the following: we are much more likely to be able to find a privileged, historical causal relation between *t* and the reference class that our intuitions assign to *t* (or, better still, the class that should be assigned to *t* given the needs of the relevant empirical theory--see note 25) if the presence of members of the assumed reference class is, for humans, an integral part of the acquisition of *t*. Under such circumstances, objections to an actual history-based approach would seem harder to come by than has often been thought. For in offering objections, we cannot proceed simply by saying, "What if the subject tokened *t* without having the right history, without, for example, having ever encountered a member of class c?" In order for such an objection to be relevant, the proposed counterfactual situation must be consistent

with the empirical theory(ies) that describes how LOT terms are individuated and acquired. And in this case, the relevant empirical theory, DS theory, seems to place substantial constraints on the range of possible causal histories consistent with the acquisition of *t*.

V. The Nature of the Constraint

In this section and the one that follows, I attempt to make clear the effect of demanding that a naturalistic semantic theory for LOT be firmly grounded on a theory of LOT term individuation. In the present section, I outline a general constraint on thought experiments. Section VI illustrates the way this constraint can bear on our theorizing by applying the constraint to a specific theory of intentional content currently on offer, Jerry Fodor's asymmetric dependence theory.

Consider a simple, naturalistic theory of reference for LOT terms, which following Fodor I will call the 'Crude Causal Theory' (Fodor 1987, p. 99). The Crude Causal Theory says that an LOT term *t* refers to whatever causes the tokening of *t* in the subject in question. Difficulties with this theory are obvious. Take the LOT term 'horse': In the typical subject, the perception of a horse causes the tokening of an LOT term that we can (by stipulation) label 'horse'; however, many other things can cause the tokening of 'horse', for example, saddles or cows on dark nights (although we can't be sure of this *a priori*--see notes 3 and 7). Some of these causes are members of kinds other than the kind *horse*. How are we to limit the extension of 'horse' to just the horses, excluding the other causes of 'horse' tokenings? Generally speaking, a naturalistic semantics needs to locate a privileged relation R that can be characterized nonsemantically and that holds between a given LOT term *t* and the individual or members of the kind that constitute the extension of *t*.[20]

In our quest for a suitable naturalistic relation R with which to replace the one identified by the Crude Causal Theory, it would seem appropriate to appeal to counterfactual considerations. Accepted philosophical method suggests that we first calculate the extension a proposed R assigns to *t* under some imagined circumstances and then check to see whether that assignment

matches our intuitive assignment of extension to *t* in those circumstances. If, in the imagined circumstances, our candidate R assigns to *t* an extension that conflicts with our intuitive verdict, then the conflict is taken to be a (possibly decisive) strike against the proposed R.

Here I give no general argument against the method of testing philosophical theses by comparing their implications to our intuitive responses to counterfactual cases; however, naturalistic methodology imposes certain constraints on this approach as it might be applied to test a semantic theory. With respect to the evaluation of a proposed R, we are commonly asked to imagine that a subject S tokens LOT term *t* under such and such conditions without being given any reason to believe that it is nomologically possible for a human to token *t* under the conditions described. Typically either a critic or proponent of a proposed R points to an LOT term *t* and tells us to imagine S tokening that very term under such and such conditions; and upon her doing so, we are asked what intuitive assignment of content we would make under those conditions and whether R makes that assignment. For the reasons set out in section II above, we should be wary of conclusions based on such acts of imagining and our intuition-based responses to them. If my explication of naturalistic methodology is correct, we are obliged, first and foremost to develop a theory of content that fits the actual facts of the human case; the theory, so developed, would then be tentatively extended to unusual cases, nonhuman species, etc., with an eye on what similarities might exist between the nonstandard cases and what has been identified as theoretically important in the human case (although there is, of course, room for give and take here, for so-called reflective equilibrium). At least in the early stages of theory development, and the stage we're in now would seem to count as such, our highest priority is to identify the ground of reference for mental representations in humans, not in some other beings who are hypothesized to be able to token *t* where, so far as we can tell, it is impossible for a human to do so (or incredibly unlikely that a human would do so--for further discussion of this caveat, see section VII). Say that we construct our best theories of human cognition, and they explain human cognitive skills by invoking LOT terms individuated in a certain way; if this way of individuating terms implies that it is nomologically impossible (or

even astronomically unlikely) for a human to token LOT term *t* under the circumstances described in a thought experiment, then the thought experiment is irrelevant (or largely irrelevant) to the development and evaluation of a naturalistic semantic theory.

I have just argued that our best theory of LOT term individuation should constrain our use of thought experiments to evaluate naturalistic semantic theories; call the constraint I've described 'NatCon' (for naturalistic constraint). Perhaps it is worth inquiring briefly after the implications of three approaches, other than the DS-based one, of individuating LOT terms, to see whether there is a way to escape, or at least render slight, NatCon's force. First recall that one common way of individuating LOT terms is off limits. Given that we are considering how thought experiments might be used to either motivate or criticize a naturalistic *semantic* theory, a given *t* should not be picked out according to its content: we want to know whether the content assigned to a given term by a naturalistic relation R is the correct assignment; thus, we must have in hand an independently identifiable term whose content is an open question, a term (1) to which we can make an intuitive content assignment (or, better, an assignment motivated by the explanatory needs of a particular theory) and (2) the tokening of which we can then inspect for its participation in relation R, so that (3) we can compare the two assignments to see whether they match.

The naturalistic orientation might suggest a second tack: we identify *t* by appeal to the best current theory of mental processes as they occur in the matter of the brain. It is sometimes proposed, for example, that a given LOT term is identical (at least for a particular subject) to some specific neural structure. Though I press no complaint against this approach,[21] I wish only to note that identifying *t* with a specific type of neural structure places a severe limitation on the range of situations in which *t* can be tokened. Such situations are limited to those in which it is nomologically possible for the subject to instantiate the neurological structure in question (Such situations should also exclude cases where there is an incredibly small chance of the structure's appearing, say, as the result of chaotic neural firing; for such a structure would, I assume, lack

the causal powers and participation in relations that would incline us toward thinking of it as an LOT term). Thus, NatCon remains substantive and in effect.

As a third alternative, consider the view that mental representations are multiply realizable (which view I will abbreviate as 'MR'). This views rests on the observation that a perfectly respectable natural kind, say, *mountain*, can take many different forms with respect to its shape or size or the stuffs out of which it is made.[22] Applied to LOT terms, MR implies that a given LOT term *t* should not be identified with a specific neural structure, because different neural structures can play the role of *t* on different occasions of its tokening, even within a single subject. According to the common functionalist elaboration of MR, mental representations possess their identities in virtue of their functional roles, i.e., their connections with each other and to inputs and outputs, and thus, facts about the physical constitution of a given token of a mental representation lie beside the point. On MR, then, it would be wrong-headed to point to a specific neural structure *s*, say that it is *t*, and then go on to claim that the subject in question only tokens *t* in cases where *s* is present in the subject's brain. Furthermore, since it would seem that a mental state or representation could, at different times, be realized by different attractors in a single subject's phase space, MR might seem to imply not only the irrelevance of neuroscience to LOT term individuation, but also the irrelevance of any constraint that might emerge from the DS approach.

MR faces metaphysical problems that threaten its very coherence (For an elaboration of these problems, see the pertinent essays in Kim 1993). Setting aside general difficulties with MR, it does not seem that accepting MR will allow us to escape the force of NatCon. First off, we should, on methodological grounds, limit our claims to the multiple realizability of a mental representation to the sorts of realizing states in which we have positive reason to believe; MR was conceived of partly as a retort to type physicalism, i.e., as a way to leave open the possibility that a being without a human's physical constitution could have mental states. While it is possible that such beings exist, and that our psychological theories will eventually have to address their mental lives, our naturalistic methods tell us to interpret MR more narrowly, as the

claim that many different neural structures (or many different attractors in phase space) can instantiate the same LOT term in a human; however, this is not the same as claiming that just any old neural structure can be (or any old attractor can count as) *t* in whatever circumstances one can dream up. If talking about terms in LOT is to yield any empirical power for an MR theorist, there will have to be substantial constraints on which of the various physical structures at what times count as a given *t* in a given human.[23] Without such constraints, any physical structure could be *t* at any time, rendering vacuous any explanations that invoke the tokening of *t*. For what explanatory power might the tokening of *t* have if, in a single subject at a single time, any neural structure can count as a tokening of any LOT term? So long as one hypothesized the right combination of terms in the subject's mind, *t* could always be "present" to explain current behavior, as would any other LOT term one happened to find convenient for explaining the subject's behavior. The MR theorist might attempt to secure the explanatory value of talk about *t*'s occurrence by citing those factors that are often thought by functionalists to individuate state types or determine content, factors such as relations to inputs and outputs. However, note the substantial nature of input- and output-based constraints on which physical structures can count as *t* for the functionalist (In the case of the mental state type pain, for example, Putnam suggests the possession of inputs that signal "damage to the Machine's body" as an input-based constraint on what can count as pain [Putnam 1967, p. 227]). In an individual subject, and perhaps in the species as a whole, there are physical limitations on which neural structures can play the role of terms for immediate sensory input (such as those that could signal damage at the periphery); there are also physiological constraints on which neural structures can be used to give motor commands; thus, even acknowledging the great flexibility and adaptability of nervous system, the physical constitution and organization of our sensory and muscular systems place substantive constraints on what can count as an LOT term for a given input or output (Physiological constraints may extend to nonsensory or motor concepts as well--see Lewis 1972, 1980). I am well aware that the functionalist includes internal relations between states and terms as part of her characterization of mental state types, and presumably, LOT term types, but there must

always be some connection at the periphery to LOT terms for inputs and outputs. Given the physically-based limitations on what can count as input and output terms, there would seem to be a limited number of physical states that could play the appropriate mediating roles *vis-à-vis* LOT terms for inputs and outputs. For these reasons, acceptance of MR does not liberate the semantic theorist to the extent that she is free to stipulate the tokening of *t* in subject S under whatever circumstances seem interesting or convenient for the purposes of constructing test cases for naturalistic semantic theories; NatCon remains in effect.

Whichever way we individuate LOT terms, once we've done this in a naturalistically-respectable (i.e., nonintentional, nonsemantic) fashion,[24] we've introduced a substantive constraint on what will count as a relevant thought experiment for the purposes of evaluating a naturalistic semantics for LOT terms. NatCon implies that if we have appealed to a particular naturalistic theory T in our individuation of LOT terms (say, DS theory), then we are obliged to look to T to inform us as to the conditions under which a given LOT term can or cannot be tokened.[25]

VI. An Example

In this section, I illustrate one way in which NatCon might bear on the evaluation of naturalistic theories of intentional content. Consider Fodor's much discussed asymmetric dependence theory of content (ADT, hereafter) (Fodor 1987, 1990a). Granted Fodor sometimes distances ADT from any claims about *human* psychology; however, if the naturalistic methodology advocated in section II is sound, then Fodor should be more interested in the way his theory applies to humans as representational systems.[26] Note also that sometimes, e.g., when discussing supporting examples and potential counterexamples, Fodor considers humans as test cases (see the various examples he considers in Fodor 1990a and Loewer and Rey 1991). That Fodor, as well as his critics (e.g., Cram 1992, Manfredi and Summerfield 1992, Seager 1993), find it hard to discuss ADT without using humans as test cases seems to bolster the reasoning of section II of the present work; for it seems to show how difficult it is to argue in a persuasive manner about the

conditions for representation while attempting to abstract completely from the human case. Regardless, once we apply ADT to the human case, for whatever reason, the abstract, hypothetical situations that Fodor describes as a means of explaining ADT run a serious risk of violating NatCon.

Take the LOT term ‘horse’ as an example. Sometimes--on dark nights, for instance--cows have the property of being causes of ‘horse’-tokenings (Fodor 1990a, pp. 121-122). While the Crude Causal Theory would automatically, and incorrectly, include cows in the extension of ‘horse’, ADT is supposed to give us a way to avoid this undesirable result. ADT’s central principle says that the LOT term ‘horse’ refers to horses rather than cows if (a) in possible worlds where the nomic connection between being a horse and being a cause of ‘horse’ tokenings has been broken (all else remaining the same as in the actual world), cows don’t cause ‘horse’, but (b) in possible worlds where the nomic connection between being a cow and being a cause of ‘horse’ tokenings is severed (all else remaining the same as in the actual world), horses still cause ‘horse’ tokenings. The second half of the principle says that we are to imagine what would happen if the property of being-a-cause-of-'horse'-tokens were nomically dissociated from cows. Upon such dissociation, we are to ask ourselves what other items, if any, would still cause the subject in question to token 'horse'. If horses still would, then the second half of the asymmetric dependence condition is satisfied, but how are we to know what will happen once the pertinent nomic alterations are made?

Understood as a theory of human LOT term content, ADT’s content assignments depend on claims about what would happen in a human cognitive system, described in nonintentional terms, in possible worlds where the causal regularities that normally hold between the system and its environment have been altered. In order to change the nomic structure of the world in the ways ADT prescribes, so that cows (even on dark nights) no longer cause the subject in question to token 'horse', it would seem that specific changes would have to be made in the nomic structure of the world: some observable properties of cows, most likely their properties of being big and four-legged, would have to be nomically dissociated from 'horse'. Thus we see that in order to

spell out ADT as a theory of intentional content for humans, Fodor must underwrite the theory with some nonintentional account of LOT term individuation. Without having given the details of such an account, Fodor runs a greater risk of violating NatCon; for we should be concerned that without a detailed theory of LOT term individuation, the effects of the suggested changes in the nomic structure of the universe are unknown to us. If an animal's being big and four-legged were to no longer cause the tokening of 'horse' in me, what concomitant changes would occur? How am I to tell? One may well think that if the properties of being big and four-legged were to no longer cause the tokening of 'horse' in me, then horses would no longer cause the tokening of 'horse' either (a serious problem for ADT). After all, horses' being big and four-legged is a very important part of my recognizing them as horses. I don't actually know what would happen if being big and four-legged were to no longer cause me to token 'horse', but it is this ignorance that constitutes my complaint: in the absence of a theory of how LOT terms are individuated, introduced, and retokened, we simply don't know what will happen in the counterfactual situations ADT directs that we imagine.

What theory might it be that Fodor has in mind, the machinery of which can be used to individuate terms in LOT? Fodor's picture combines a theory at the neural level with one at the computational, formal, or symbolic level. "Tokens of symbols are physical particulars in good standing. I suppose this to be true, *inter alia*, of tokens of mental representations which are, presumably, neural objects." (Fodor 1990b, p. 315)[27] According to this now familiar picture, a human's system of neural structures instantiates a system of symbols and rules. Such rules are algorithms, perhaps represented only implicitly, computed by way of sensitivity to nonintentional, nonsemantic properties of the symbols. Fodor's views suggest two approaches to nonsemantic term individuation, both considered in Section V: a symbol is of a given type in virtue of its possessing certain neurophysiological properties; or, a symbol is of a given type in virtue of its functional/computational role. Either theory Fodor chooses substantively constrains claims regarding term tokening under counterfactual conditions; this was the primary conclusion

of Section V. Again, my concern is not so much that ADT is false, but that we cannot confidently evaluate ADT lacking as it is in requisite detail.

Numerous complications would need to be addressed if we (or Fodor) were to try to fill in a picture of LOT term individuation in anything that approaches sufficient detail. For example, were we to decide on a neurophysiological approach to LOT term individuation and were we to locate a subject in whose current tokening of 'horse' we are confident, we would still have to isolate in the subject's overall neural profile that portion which is to be identified as 'horse'. It is not my intent here to spell out such details; however, the viability of ADT as a theory of content for human mental representation depends on those details. In order to make informed judgements as to which LOT terms will be tokened in nomically altered worlds, we must have a clear idea of the form those terms take in the brain/cognitive system.

Once we have characterized a particular relation R (and ADT stands as one possible characterization), someone can raise an objection of the following form: "I can think of a case where R holds between *t* and some collection c, but where c is not what we would intuitively take to be the extension of *t*." Call objections of this form 'CER' (for 'counterexample to R') objections. Fodor introduces ADT largely as a response to CERs pressed against the Crude Causal Theory. Fodor assumes that the Crude Causal Theory needs fixing largely because it is so easy to describe situations where the relevant R holds (cows on dark nights cause 'horse'-tokenings) but where the resulting extension assignment is incorrect (because 'horse' doesn't refer to cows, on dark nights or otherwise) (Fodor 1987, p. 101). However, before dismissing the Crude Causal Theory, or any other naturalistic theory, on the basis of a CER objection, we should bear in mind the relativization demanded by NatCon. We should not take just any logically consistent CER objection to be an effective criticism of the theory of reference built on R. When evaluating a naturalistic semantic theory, we should be generally suspicious of CER objections, at least until they're relativized to the particular empirical theory or group of theories that nonsemantically specify the individuation criteria for LOT terms (a matter about which, recall, we have very few, relevant, pretheoretical intuitions). Proper relativization changes the

CER objection schema into the following: "I can describe a case that is consistent with the relevant empirical theory(ies) of LOT term individuation and where R holds between *t* and extension class c, but where c is not the extension of *t*." (Call objections which fit this modified schema 'CER*' objections, where '*' is to be replaced in an individual case with a reference to the relevant relativizing theory(ies).) Once we've adopted the DS approach to individuating LOT terms, for example, any CER objection to a naturalistic semantic theory for mental representations must take the form of a CER-DS objection.

VII. Swampman and the Role of Causal History in the Determination of Content

No matter how much the appearance of a particular LOT term in an actual human seems to depend on the subject's causal history, many philosophers are tempted to view this consideration as a red herring in discussions of theories of intentional content. "We're interested in the metaphysical nature of content," the cry rings out, "not the embodiment of content within one particular nomic setting."

In response to this intellectual tendency, I reiterate the claim of section II: sound naturalistic methodology directs us to investigate central cases as we actually encounter them. Without entering into a full debate over the merits of the naturalistic approach, once we have seen fit to take naturalism seriously, the recent history of empirical investigations of human concepts seems to support the secondary status here assigned to *a priori* meditations on the nature of content. The empirical data on concept structure and acquisition do not seem to support the idea that humans are in touch with abstract concepts (say, Platonic forms or Fregean intensions) that precisely determine extensions across all possible worlds. The data instead suggest that human concepts are more like rough and ready identification procedures, which may serve us well in our dealings with the surrounding world, but may not determine extensions in unusual or nomically altered environments.[28] More to the point, it seems that the developing child constructs the very concepts *meaning* and *representation* expressly for the purposes of explaining and predicting the behavior of cognitive systems (including the thinker herself) in the

actual world (Astington 1993, Flavell et al. 1995, Gopnick and Wellman 1992); given the imprecision of these concepts and the limited use for which they were developed, it would seem wrong-headed to accord much theoretical weight to intuitions regarding the correct application of the concepts of meaning and representation under nonstandard, perhaps even nomologically impossible, conditions.

All of this is perfectly consistent with the existence of a well-delineated natural relation of reference, which cognitive science, with what aid it may get from pretheoretical intuitions, should investigate. This does not, however, imply that cognitive scientists should try to find, or will be able to find, a reference relation that is fully determinate across all possible worlds. From the standpoint of the naturalist, the motivation behind such a desideratum is unclear. If the goal is to provide a full analysis of our *concept* of representation, then, as I have already suggested, the naturalist has no reason to think that the concept will resolve into necessary and sufficient conditions. If, instead, the goal of locating a reference relation that is fully determinate across all possible worlds is motivated by a desire to find out what relation we're actually referring to when we talk about reference, then the methodology described in section II suggests we respond by acknowledging the legitimacy of the motivation, without thereby certifying the goal. Rather than attempting to identify necessary and sufficient conditions for reference, i.e., conditions that determine reference in all possible worlds, we should investigate the reference relation that exists in this world, starting with what seem its most obvious instances.

To make more concrete the philosophical concern to which the preceding is a response, consider a particular thought experiment, due to Donald Davidson, often thought to have important philosophical implications. Imagine a being who appears out of nowhere, in a swamp, as it were, and who acts just like a person. (According to Davidson's original description, the Swampman is a molecule for molecule replica of Davidson himself, amazingly brought into existence by a bolt of lightning that simultaneously destroys Davidson, but these details are unimportant for present purposes; see Davidson 1987, p. 443.) Whether a being that appears out of nowhere qualifies as a *person* is a greatly disputed matter (See the contributions to the

Swampman forum in *Mind and Language* 1996). However, given that, *ex hypothesi*, the behavior of Davidson's Swampman is indistinguishable from that of a human, it is thought that our theories addressing matters intentional should issue a definite and correct pronouncement regarding the nature and content, if any, of Swampman's intentional states.

For Swampman's possible appearance to bear directly on the current discussion, the possibility of Swampman should provide the basis of a CER-DS objection to the type of causal theory of reference for LOT terms for which I have been attempting to make room. I have not filled in the details of such a theory of reference, but for present purposes, there is only one trait that matters: an actual history-based causal theory assigns reference to a semantically basic LOT term as a function of past causal interaction between the subject and her environment. Swampman, however, has not interacted with his environment in any way at the time when he first appears on the scene; yet by DS standards, Swampman would seem to possess LOT terms: the appropriate phase portrait would seem to contain a pattern of attractors similar to the patterns found in phase spaces for typical human subjects. This opens the door for a CER-DS objection: to the extent that we are willing to attribute extensions to any of Swampman's LOT terms, actual history-based causal theories fail; for Swampman's terms will have reference without any causal history.

The naturalist can legitimately respond in at least the following two ways: First, given our lack of experience with anything of Swampman's nature or ilk, we should simply ignore the example, and any of our reactions to it, on methodological grounds. Given the methodology advocated in section II, there is no reason to take seriously anyone's intuitions about Swampman's representational capacities; our primary goal is to develop a naturalistically respectable theory of representation, one that (a) accommodates our reactions to the real representational systems with which we have had actual experience and (b) explains the behavior of such actual systems. Given this goal, we should not accept intuitions about a hypothetical Swampman as data points of sufficient weight to motivate *or count against* an otherwise well-founded naturalistic theory of representation. Recall the point made above about the nature of

humans' concepts of meaning and representation: given that these concepts develop, from an early age, as ways to explain the behavior of actual representational systems, we should expect that many persons will lack firm intuitions about the content of Swampman's internal states; we should also expect the naturalist to be suspicious, and rightly so, of the intuitions of those who claim to have reliable, pretheoretical insight into the nature of Swampman's mental life.

Secondly, the naturalist should be concerned about the discussion of Swampman even as a hypothetical being: so far as we know, it is impossible for a fully formed representational system to appear out of nowhere; such spontaneous generation goes so much against what we know about the genesis of representational systems that its alleged possibility seems irrelevant from the naturalist's standpoint. Accordingly, the naturalistic philosopher of psychology is no more obliged to provide a theory of content applicable to Swampman than is the nuclear chemist obliged to say where on the periodic table would fall an element each of whose atoms consists of 4,000 protons, 1 neutron, and 5 million electrons (cf. Dennett 1996, 76-7). To defend this approach to the Swampman example, the naturalist would seem to need a more detailed view of what is possible, and more to the point, which possibilities count when evaluating a theoretical proposal in the natural sciences. It may be that both Swampman and my odd atom, though incredibly unlikely to appear, are nomically possible in some sense (i.e., their appearances don't violate fundamental equations in quantum physics or general relativity); however, it seems that there is a significant sense in which they are methodologically irrelevant, and we would like some theoretical apparatus to explain why. Though I will not attempt to elaborate any such theory here, the actual practice of natural scientists should serve as its basis: we should be quite surprised, I think, to find working scientists refusing an otherwise perfectly good theory on the ground that it is inconsistent with an unactualized possibility that is incredibly unlikely to ever be actualized (This is especially so in a case such as Swampman's, where, even if the example is allowed as relevant, the force of the thought experiment depends on intuitions of which we have independent reason to be suspicious--see the naturalist's first response given above). Faced with this type of situation, the physicist's response seems to be to accept the otherwise perfectly good

theory, but renormalize when necessary. Similarly, if we can construct a perfectly good theory of reference based on our attention to the actual histories of cognitive systems, we should stand firm against those who claim that the unactualized possibility of Swampman refutes our theory. Part of what is at issue here is the burden of proof. My claim is that from an examination of scientific practice, the following guideline should emerge: when the naturalist is asked to take an incredibly far-fetched possibility seriously, as providing a possible refutation of an otherwise well-motivated theory, the burden of proof is on the person doing the asking; and in the case at hand, this means that she must show there to be a significant possibility that Swampman will appear and persist, and she must do so in a more convincing manner than by simply saying that Swampman's appearance and persistence is not impossible. Natural science's job is to explain actual phenomena, not merely possible ones; and so it goes for a naturalistic theory of reference for LOT terms.

These blanket naturalistic responses to Swampman-related worries may seem a bit extreme. Given that Swampman is described by Davidson to be a molecule for molecule duplicate of Davidson himself, we should be able to say something sensible about the meaning, if any, of Swampman's thoughts, should we not? Surely we don't want to rule out the possibility that Swampman has mental states (or can token mental representations) solely on the grounds that he (it?) is not a normal human. Here is a concession I think that the naturalist should be willing to make. Insofar as we have in hand empirically successful theories of content developed on the basis of our study of human capacities, we should be willing to apply these theories to Swampman by analogy (cf. Neander 1996, p.127). For example, we may want to say (hypothetically, of course) that Swampman has intentional states with a certain content because his (hypothetical) phase space is in a shape that is exactly like ones that we know to have been shaped by a process that allows for reference-fixing. The point here is that while we should be able to make some sense of Swampman's "behavior", the Swampman example should not be the tail that wags the theoretical canine:[29] intuitively decided upon, "right" responses to philosophical dilemmas posed by Swampman's hypothetical existence should not be used to

drive an application of modus tollens meant to disprove a naturalistically sound theory of content. If Swampman has intentional states at all, it's because he has something comparable to the content identified by a naturalistically sound theory of content. Even this suggests too much of a concession. Returning to the points made above, if Swampman were to exist, he wouldn't be Swampman, for he would possess a real history. After all, so far as we can tell, there are no such things as swamp-persons.

What bearing, then, does the naturalistic defense against swamp-people have on our evaluation of causal history-based theories of reference for LOT terms? Can the critic not put aside fantastical swamp-people and still formulate fundamental, in-principle objections to actual history-based theories of reference? What if, to take another commonly discussed type of example, a super scientist raises a human from *in vitro* fertilization on through childhood, the whole time manipulating the input to this human's sensory systems (and whatever other systems necessary) so that the human acquires a wide range of LOT terms under conditions that would lead our favored, actual history-based theory of reference, whatever it might be, to assign the wrong extensions to the captive human's LOT terms? Here the objector has removed the mystery of a person who has no learning history at all and replaced it with the intuitively more plausible example of a person whose input has been freakishly manipulated to create a system best modelled by a standard phase portrait, despite a nonstandard history.

Given existing, quite appropriate ethical standards for research using human subjects, it would be unfair for me to respond to the critic by challenging her to play super scientist and prove that the example, as described, is nomologically possible. But as a condition for our taking the super scientist example seriously, it is fair to shift the burden of proof to the critic: we should require her to show that the super scientist story coheres, *prima facie*, with the theories in cognitive science to which our theory of content is to be relativized. If our best theory of human LOT term acquisition and individuation, DS theory, as I've suggested, is at odds with the super scientist story, and it would seem to be, then we can dismiss the super scientist story as a violation of the constraint built into the CER* objection schema.

Recall the distinction introduced in section IV between LOT terms whose extensions are fixed independently of the content of other mental representations and those LOT terms whose extensions are not so fixed. Bearing in mind that our present focus is on terms of the former type, the semantically basic terms, we can see more easily why the super scientist story does not meet automatic success as a counterexample to an actual history-based theory of reference for LOT terms. Consider again the infant's LOT term 'object' as an example of a semantically basic LOT term. Some recognizable version of the object concept is present in the infant from a very early age and would seem to be a fundamental term in the infant's LOT. How in the world will the super scientist shape the infant's phase space, causing the infant to acquire 'object', without exposing the infant to objects as stimuli (thus introducing the possibility of the relevant R's holding between 'object' and objects)? It's easy to say that the super scientist will just 'stimulate the right neurons'. However, if Thelen is correct in claiming that experiences in the development of motor control are the rudiments of cognition, then the super scientist has her work cut out for her. The challenge becomes even more daunting given that it is very likely that the set of differential equations that fully describes the workings of the human cognitive system is nonlinear and, at least by present lights, cannot be solved. (This is part of the reason why much of the research into human cognition that applies the tools of dynamical systems theory depends on qualitative modelling, rather than quantitatively precise modelling.) The scientist would have to be quite super after all, perhaps to the point of being a nomological impossibility herself. More likely than not, then, the only way for a plausibly super scientist to get the infant to master the LOT term 'object' is for the scientist to do what other parents do, i.e., show the infant some objects. This, of course, would establish a historical, causal connection between objects and the infant's LOT term 'object', which can be exploited by an actual history-based theory of reference for LOT terms to explain how 'object' achieves its correct extension, the collection of objects.

VIII. Objections: On the Incompatibility of LOT and the DS Approach

In this section, I address two objections to my use of the DS view of cognition: first, that the DS-based view of LOT term individuation does not allow for meaningful interpersonal (possibly even intrapersonal) comparisons, and second, that according to the DS-based view of cognitive processing, attractors don't behave enough like repeatable, linguistic units for us to reasonably think of them as terms in LOT.[30]

In more detail, the first objection runs as follows: Assume that the DS theory of cognition is correct, and, in particular, that a subject's interactions with her environment significantly shape her developing cognitive system in the way described by DS theorists working in the area of developmental psychology. Given the large degree of variation in subjects' individual histories, no two persons (perhaps not even one person at different points in her own life) will share the same cognitive profile as this is characterized by the appropriate phase portrait; the set of attractors will differ from one subject to the next (cf. Thelen 1995, pp. 86-90). Thus, given that I've identified LOT terms with attractors, there seems to be no point in talking about the comparative content of elements two different subjects' LOT terms.

I believe I can assuage this first concern by reiterating what I take to be the relevance of DS theory to my argument in this paper. By invoking the DS view, I intend to characterize LOT terms nonsemantically, i.e., by no mention of their reference or content; we are to attribute reference to such terms on independent grounds. I also claim that given the way interaction with the environment shapes attractors, a theory that assigns reference on the basis of past causal interactions has a reasonable chance at success, this because it seems possible that as interaction molds the attractors, reference-fixing causal relations are given a toehold; at the very least, the need for interaction with the environment as the route to attractor emergence secures the opportunity for reference-fixing causal relations to enter the picture. As I see things, no problem results from the fact that abstract mathematical profiles differ from one subject to the next; all that matters for the purpose of making meaningful comparisons across subjects is that each

subject possess a term with the relevant referential properties. Note, however, that possessing terms with the *same reference* does not require that two subjects possess the *same attractor*: Assume that subject *a*'s experience with objects shaped an attractor in *a*; and because the appropriate reference-fixing causal interactions occurred between the developing attractor and objects, the attractor in question refers to objects. Similarly for another subject, *b*: assume that in *b*, there exists an attractor that emerged as a result of *b*'s interaction with objects; further, assume that enough of this interaction was of the right sort to have fixed the extension of *b*'s attractor as the collection of objects in the world. This in no way implies that *a*'s and *b*'s relevant attractors are identical. From the standpoint of a naturalistic theory of reference for LOT terms, and any accompanying externalist psychological explanations, all that matters is that the two attractors have the same reference, different as those attractors might be.

This situation as I've described it can be roughly, but I think instructively, compared to the situation as it stands in the linguistic context, where, for example, 'cat' refers to cats, and so does 'gato'. From the semanticist's standpoint, it is irrelevant that the words are different in form; at least if her interest is in referential semantics, as I am concerned with here, the linguist will care only that the two words refer to the same type of animal. Furthermore, note that reference might have been fixed for both terms by the same general reference fixing process. In the historical development of each language, we can imagine the occurrence of similar types of events by which the reference of the terms was fixed: the appropriate speakers pointed to extant cats and said "'cat' will refer to all things of the same natural kind as those animals over there" (substituting Spanish words throughout when we turn to 'gato') (Putnam 1975). My point is simply this: when we construct theories of reference for natural languages, we worry not that the forms of words may differ from language to language; so long as the relevant terms refer to the same thing or group of things, we can make at least that meaningful comparison--and from the standpoint of the philosopher developing a naturalistic theory of reference for LOT terms, the reference-related comparison is the one that counts.[31]

The second objection charts a course to controversial territory, to the lively debate over constituency and compositionality in connectionist networks, and now, dynamical cognitive systems.[32] I can do little to resolve the debate here, but we should keep in mind how the worry about constituent structure relates to the view I have put forward. Some parties to the debate over constituent structure claim that attractors corresponding to individual terms do not combine to yield composite representations in the way terms from a public language do: a system of representation by attractors has neither a combinatorial syntax nor a compositional semantics. This seems to introduce a gap between the DS view and any talk of LOT: It is the very idea that the medium of thought possesses a combinatorial syntax and a compositional semantics that inspires the idea of a *language* of thought in the first place; if these comparisons between natural language and the mental medium degenerate, then talk of LOT seems out place; for LOT no longer possesses what is supposed to be distinctive of it. Fundamental here is the claim that an attractor representing a mental state as a whole, e.g., a belief, does not consist of a physical concatenation of atomic parts; the lack of identifiable constituents stands at odds with the idea of combinatorial syntax and compositional semantics for the following reasons: If a language is governed by a combinatorial syntax, well-formedness in the language should be defined by (perhaps implicit) syntactic rules; such rules state legal patterns of combinations of atomic elements and do so by referring to the syntactic categories, noun and verb, for instance, to which the atomic elements belong (allowances being made here for some reference to syncategorematic elements). 'John loves Mary' is a legal string in English, and it is so because, roughly speaking, 'loves' is a transitive verb flanked by the nouns 'John' and 'Mary'; but if the sentence as a whole were to lack the identifiable nominal element 'John' (as an attractor representing the complete thought is said to), the explanation of the well-formedness of the whole sentence would seem to make no sense: such an explanation cannot appeal to the fact that the verb is preceded by a noun, when there is no noun there to be found. Similar remarks apply to compositional semantics: A scheme of representation has a compositional semantics if and only if the semantic value of an entire legal string of that system is a function of the semantic values of the string's constituent

parts. If the entire thought ‘John loves Mary’ contains no identifiable element ‘John’ to whose semantic value we can appeal when calculating the semantic value of the entire sentence, it is unclear how a compositional semantics can apply to ‘John loves Mary’. Therefore, if attractors representing complete mental states lack constituent structure, it would seem pointless for us to identify certain attractors as terms in LOT and go on to assign reference to them; attractors corresponding to individual terms do not appear as parts of complete representations, and thus the reference we assign to them (as well as their assigned syntactic properties) remains inert: if the individual terms do not appear as parts of the whole mental state, they cannot contribute their individual referential meanings to the representational content of the complete thought (and, likewise, their alleged syntactic properties can have no effect on the subject’s thought processes).

What seems to be missing from my account is an explanation of how attractors combine, so that it will make sense to say that the subject’s tokening of ‘John loves Mary’ in LOT consists partly in the tokening of an attractor identical to ‘John’. Here I can do little more than point elsewhere, but I will at least suggest some directions the DS theorist might take in the attempt to identify LOT terms as recurring constituents of larger cognitive structures. As a beginning, consider Elman’s use of principal components analysis (PCA) to identify privileged areas of the phase space (Elman 1995, 210-15). When applying PCA to analyze a system’s behavior, we plot values of certain parameters of interest while suppressing others; and by doing so, we highlight patterns in the behavior of the system that we might not notice otherwise. Elman applies PCA to characterize syntactic categories, rather than to identify privileged dimensions along which certain values are to be associated with the tokening of an individual LOT term; however, PCA may also hold promise as means for achieving the latter goal. Further, Petitot applies the morphodynamical approach to describe both the attractors that represent syntactic categories (1995, pp. 233-4 and passim) and to identify individual terms (1995, pp. 250-1). Lastly, recent work by Paul Churchland (Churchland 1998) may be of some interest here; Churchland describes how one can apply Gutman point alienation, useful for performing certain statistical analyses, to meaningfully measure structural similarity across two or more of an important

subvariety of dynamical systems, connectionist networks; note that we can also apply such measures of similarity to compare the same system at different times in its history. The work I've cited is, to a great extent, in its early stages; however, given the richness and variety of such work, it does not seem overly optimistic to think that we will uncover LOT terms that reappear, though not in an immediately obvious way, as constituents of the complete mental states found in cognitive dynamical systems.[33]

The analytical methods toward which I've no more than gestured may provide the tools for individuating syntactically robust LOT terms within the structure of a DS theory of cognition. However, one might worry that if DS theorists realize this promise, they may thereby relegate DS models to the status of mere 'implementation-level models', thus rendering DS models uninteresting from the standpoint of the cognitive theorist (Horgan and Tienson 1994, p. 327); for shouldn't true cognitive theorists be interested in thought itself, rather than how cognitive activity is implemented by one particular type of cognitive being? Would it were so simple a matter. Alas, if we take naturalistic methodology seriously, we have no conceptual or *a priori* grounds for cleanly separating essentially cognitive activity from mere implementation. If, as would seem to be the case, humans are the only full-blown representational systems to which we have experimental access, we had best examine carefully how humans represent the world; and if humans develop their capacity to represent partly via the emergence of attractors in the individual's phase space, then properties possessed by these attractors cannot be dismissed as mere implementation details--at least not until we have a much more thoroughly worked out theory of cognition than we have at present.[34]

I have left a number of issues unresolved; in closing I briefly comment on two of these: An issue of great importance to many cognitive theorists is that of the causal efficacy of constituent structure, i.e., whether the constituents identified using analytical methods have causal effects on the system, qua constituents; many of the works cited in note 32 take up this issue, with some authors claiming that even if constituent structure can be uncovered in dynamical cognitive models, such structure is not causally efficacious or explanatorily relevant (Garson 1997,

Ramsey 1997). This is a dispute worthy of our interest, but deserves far more attention than present space allows.

Secondly, we might wonder about the value of assigning referents as semantic contents, the concern being that DS theorists themselves tend to endorse functional role theories of content. For example, though I have appealed to Petitot's approach as a possible means of characterizing LOT terms, Petitot seems inclined to assign content on the basis of internal relations among attractors (Petitot 1995, p. 243). Horgan and Tienson take a similar view: though they argue for an interpretation of DS models according to which syntactic properties are causally efficacious, they characterize the content of mental states internally, in terms of mathematical and structural relations among attractors (Horgan and Tienson 1996, pp. 155, 164). Paul Churchland seems to acknowledge the importance of reference-determining connections to the outside world, but on the question of representational content, he remains a two-factor theorist (Churchland 1998, p. 29). My position is that regardless of the importance of internal relations between attractors, we should take seriously and attempt to develop a theory of reference for LOT terms according to which reference is determined, at least at the basic level, by our commerce with items in the external world to which our thoughts refer. Limitations of space prevent me from defending this view here; nevertheless, it is because of what I take to be the advantages of a causal, covariational, or informational theory of reference for semantically basic LOT terms that I have appropriated the tools of DS theory in the way that I have, setting aside the internalist, functionalist approach to the determination of content that many DS theorists embrace.

REFERENCES

Armstrong, S. L., Gleitman, L. R., and Gleitman, H.: 1983, 'What Some Concepts Might Not Be', *Cognition* **13**, 263-308.

Astington, J. W.: 1993, *The Child's Discovery of the Mind*, Harvard University Press, Cambridge, MA.

Baker, L. R.: 1993, 'Metaphysics and Mental Causation', in J. Heil and A. Mele (eds.), *Mental Causation*, Oxford University Press, Oxford.

Bartsch, R.: 1996, 'The Relationship Between Connectionist Models and a Dynamic Data-Oriented Theory of Concept Formation', *Synthese* **108**, 421-54.

Block, N. (ed.): 1980, *Readings in the Philosophy of Psychology, Volume One*, Harvard University Press, Harvard, MA.

Bower, T. G. R.: 1989, *The Rational Infant: Learning in Infancy*, W. H. Freeman and Company, New York.

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

Churchland, P. M.: 1998, 'Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered', *Journal of Philosophy* **95**, 5-32.

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

Clark, A.: 1991, 'In Defense of Explicit Rules', in W. Ramsay, S. P. Stich, and D. E. Rumelhart (eds.), *Philosophy and Connectionist Theory*, Lawrence Earlbaum Associates, Hillsdale, NJ.

Clark, A.: 1995, 'Moving Minds: Situating Content in the Service of Real-time Success', in J. Tomberlin (ed.), *Philosophical Perspectives, 9: AI, Connectionism and Philosophical Psychology*, Ridgeview, Atascadero, CA, 89-104.

Clark, A. and Toribio, J.: 1994, 'Doing without Representing?', *Synthese* **101**, 401-31.

Cram, H.: 1992, 'Fodor's Causal Theory of Representation', *The Philosophical Quarterly* **42**, 56-70.

Cummins, R.: 1986, 'Inexplicit Information', in M. Brand and R. M. Harnish (eds.), *The*

Representation of Knowledge and Belief, University of Arizona Press, Tucson, AZ, 116-26.

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

Cummins, R.: 1991, 'The Role of Mental Meaning in Psychological Explanation', in B. McLaughlin (ed.), *Dretske and His Critics*, Blackwell, Oxford, 102-17.

Davidson, D.: 1987, 'Knowing One's Own Mind', in the *Proceedings and Addresses of the American Philosophical Association* **60**, 441-58.

Dretske, F.: 1981, *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA.

Dretske, F.: 1988, *Explaining Behavior*, MIT Press, Cambridge, MA.

Dennett, D.: 1996, 'Cow-Sharks, Magnets, and Swampman', *Mind and Language* **11**, 76-7.

Devitt, M.: 1994, 'The Methodology of Naturalistic Semantics', *Journal of Philosophy* **91**, 545-72.

Elman, J.: 1995, 'Language as a Dynamical System', in Port and Van Gelder 1995, 195-225.

Flavell, J. H., Green, F. L., and Flavell, E. R.: 1995, *Young Children's Knowledge about Thinking*, University of Chicago Press, Chicago, IL.

Fodor, J.: 1974, 'Special Sciences', *Synthese* **28**, 77-115.

Fodor, J.: 1975, *The Language of Thought*, Harvard University Press, Cambridge, MA.

Fodor, J.: 1981, *RePresentations*, MIT Press, Cambridge, MA.

Fodor, J.: 1987, *Psychosemantics*, MIT Press, Cambridge, MA.

Fodor, J.: 1990a, *A Theory of Content and Other Essays*, MIT Press, Cambridge, MA.

Fodor, J.: 1990b, 'Psychosemantics or: Where do Truth Conditions Come From?', in W. G. Lycan (ed.), *Mind and Cognition: A Reader*, Blackwell, Oxford, 312-37.

Fodor, J: 1994, *The Elm and the Expert*, MIT Press, Cambridge, MA.

Fodor, J., and Lepore, E.: 1992, *Holism: A Shopper's Guide*, Blackwell, Oxford.

Fodor, J., and Lepore, E.: 1996, 'Paul Churchland and State Space Semantics', in R. N. McCauley (ed.), *The Churchlands and Their Critics*, Blackwell, Oxford, 145-59.

Fodor, J., and Pylyshyn, Z.: 1988, 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition* **28**, 3-71.

Garson, J.: 1997, 'Syntax in a Dynamic Brain', *Synthese* **110**, 343-55.

Giunti, M.: 1995, 'Dynamical Models of Cognition', in Port and Van Gelder 1995, 549-71.

Gopnick, A. and Wellman, H. M.: 1992, 'Why the Child's Theory of Mind Really Is a Theory', *Mind and Language* **7**, 145-71.

Horgan, T., and Tienson, J.: 1994, 'A Nonclassical Framework for Cognitive Science', *Synthese* **101**, 305-45.

Horgan, T., and Tienson, J.: 1996, *Connectionism and the Philosophy of Psychology*, MIT Press, Cambridge, MA.

Karmiloff-Smith, A.: 1992, *Beyond Modularity: A Developmental Perspective on Cognitive Science*, MIT Press, Cambridge, MA.

Keil, F.: 1989, *Concepts, Kinds, and Cognitive Development*, MIT Press, Cambridge, MA.

Keil, F.: 1990, 'Constraints on Constraints: Surveying the Epigenetic Landscape', *Cognitive Science* **14**, 135-68.

Kim, J.: 1993, *Supervenience and Mind*, Cambridge University Press, Cambridge, UK.

Lewis, D.: 1972, 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy* **50**, 249-58. Reprinted in Block 1980.

Lewis, D.: 1980, 'Mad Pain and Martian Pain', in Block 1980, 216-22.

Loewer, B and Rey, G (eds.): 1991, *Meaning in Mind: Fodor and his Critics*, Blackwell, Oxford.

Macdonald, C., and Macdonald, G.: 1995, *Connectionism: Debates on Psychological Explanation*, Blackwell, Oxford.

Maloney, J. C.: 1994, 'Content: Covariation, Control, and Contingency', *Synthese* **100**, 241-90.

Manfredi, P. A., and Summerfield, D. M.: 1992, 'Robustness without Asymmetry: A Flaw in Fodor's Theory of Content', *Philosophical Studies* **66**, 261-83.

McClamrock, R.: 1995, *Existential Cognition*, University of Chicago Press, Chicago, IL.

Murphy, G. L.: 1988, 'Comprehending Complex Concepts', *Cognitive Science* **12**, 529-62.

Murphy, G. L., and Medin, D. L.: 1985, 'The Role of Theories in Conceptual Coherence'. *Psychological Review* **92**, 289-316.

Neander, K.: 1996, ‘Swampman Meets Swampcow’, *Mind and Language* **11**, 118-29

Norton, A.: 1995, ‘Dynamics: An Introduction’, in Port and Van Gelder 1995, 45-68.

Pessin, A.: 1995, ‘Mentalese Syntax: Between a Rock and Two Hard Places’, *Philosophical Studies* **78**, 33-53.

Petitot, J.: 1995, ‘Morphodynamics and Attractor Syntax: Constituency in Visual Perception and Cognitive Grammar’, in Port and Van Gelder 1995, 227-81.

Port, R. F., and Van Gelder, T. (eds.): 1995, *Mind as Motion*, MIT Press, Cambridge, MA.

Putnam, H.: 1967, ‘Psychological Predicates’, in W. H. Capitan and D. D. Merrill (eds.), *Art, Mind, and Religion*, University of Pittsburgh Press, Pittsburgh, PA, 37-48. Reprinted as ‘The Nature of Mental States’ in Block 1980, 223-31. Page references are to the Block volume.

Putnam, H.: 1975, *Mind, Language, and Reality*, Cambridge University Press, Cambridge, UK.

Ramsey, W.: 1997, ‘Do Connectionist Representations Earn Their Explanatory Keep?’, *Mind and Language* **12**, 34-66.

Rockwell, W. T.: 1994, ‘On What the Mind Is Identical With’, *Philosophical Psychology* **7**, 307-23.

Rupert, R. D.: forthcoming, 'The Best Test Theory of Extension: First Principle(s)', *Mind and Language*.

Seager, W.: 1993, ‘Fodor’s Theory of Content: Problems and Objections’, *Philosophy of Science* **60**, 262-77.

Shapiro, L.: 1996, ‘Representation from Bottom and Top’, *Canadian Journal of Philosophy* **26**, 523-42.

Smith, E. E.: 1989, 'Concepts and Induction', in M. I. Posner (ed.), *Foundations of Cognitive Science*, MIT Press, Cambridge, MA, 501-26.

Smith, E., and Medin, D.: 1981, *Categories and Concepts*, Harvard University Press, Cambridge, MA.

Smolensky, P.: 1991, ‘Connectionism, Constituency, and the Language of Thought’, in Loewer and Rey 1991, 201-27.

Spelke, E. S.: 1990, ‘Origins of Visual Knowledge’, in D. Osherson, S. Kosslyn, and J. Hollerbach (eds.), *Visual Cognition and Action: An Invitation to Cognitive Science, Vol. 2*,

MIT Press, Cambridge, MA, 99-127.

Spelke, E. S.: 1991, 'Physical Knowledge in Infancy: Reflections on Piaget's Theory', in S. Carey and R. Gelman (eds.), *The Epigenesis of Mind*, Lawrence Erlbaum Associates, Hillsdale, NJ, 133-69.

Stich, S.: 1983, *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, MA.

Stillings, N. A., Feinstein, M. H., Garfield, J. L., Rissland, E. L., Rosenbaum, D. A., Weisler, S. E., and Baker-Ward, L.: 1987, *Cognitive Science: An Introduction*, MIT Press, Cambridge, MA.

Thelen, E.: 1995, 'Time-Scale Dynamics and the Development of an Embodied Cognition', in Port and Van Gelder 1995, 69-100.

Van Gelder, T.: 1990, 'Compositionality: A Connectionist Variation on a Classical Theme', *Cognitive Science* **14**, 355-84.

Van Gelder, T., and Port, R. F.: 1995, 'It's About Time: An Overview of the Dynamical Approach to Cognition', in Port and Van Gelder 1995, 1-43.

Wittgenstein, L.: 1953, *Philosophical Investigations*, translated by G. E. M. Anscombe, Macmillan, New York.

NOTES

* A shorter version of this paper was presented to the 1997 meeting of the New Mexico and West Texas Philosophical Society; my thanks to the audience for stimulating questions and observations. I would also like to thank Mariam Thalos and anonymous referees for helpful comments on earlier drafts of this paper.

---

[1] Although I often refer to these representations as 'terms in LOT', the arguments presented in this paper do not imply the truth of the full-blown language of thought hypothesis as advanced by Jerry Fodor (Fodor 1975 and 1987, Appendix).

[2] Here I do not attempt to develop a theory of content. Thus, I am not concerned with the subtle differences between the meanings of the following terms: 'extension', 'reference', 'referential content', 'intentional content', 'representational content', and 'extensional content'. While the shades of meaning may vary, all of these terms are commonly used to describe the semantic content of representations, in the sense that each of the terms is used to denote either a mind-world 'refers to' relation or a type of content that presumes such a relation. Generally, I use these terms interchangeably; however, where context clearly demands the use of one over the others, I am careful to employ that term.

[3] What's said here may suggest that we strike off in search of a brute physical description of a given LOT term *t* and having succeeded at this, we then look to discover a nonsemantic relation in virtue of which *t* refers. This is misleading as a description of how we might actually discover the physical forms *t* takes. In scientific practice, we will surely have to look beyond descriptions couched in terms of the relevant physical sciences. For example, we may have to use intuitions as to when a term with the same *content* is likely to be tokened to help us identify the range of physical variation that tokens of *t* can exhibit, the assumption being that if the same content is being represented then the subject is likely to be tokening the same term. This would seem to work especially well when we can present stimuli to the subject under conditions such that the stimuli are very likely to cause thoughts with a certain content, for example, in cases where thought content is directly related to immediate sensory observations

(See Fodor 1990b, especially note 10). Here we reason from the principle that similar causes at the sensory periphery will lead to similar effects, in terms of what content is represented as well as in terms of what vehicle that content rides. Reasoning from like effect to like cause, we might pursue a similar strategy with respect to motor commands; if the subject repeatedly engages in movement *x*, and *x* is always preceded by neural activity *y*, we might infer that *y* is a repeated LOT term with content 'do *x*'. Though there are some obvious practical advantages to this bootstrapping method, by which we attribute contents in order to identify the vehicles of content, note that (a) the vehicles must ultimately have content-independent individuation criteria (cf. Pessin 1995, pp. 34-5, 41-2) and (b) from a practical standpoint, we must have some idea what the individuation criteria are before we begin the bootstrapping process; for we need to know what sort of common element (certain types of patterns of neural activity, for example) to look for in the physical profile of the subject when we think that she has tokened a term with the same content on two different occasions.

[4] While few theories on offer make content a simple function of the subject's past history, the theories of content offered in Dretske 1981, Dretske 1988, and Maloney 1994 assign an essential role to the individual subject's past history (learning history, in particular) in the fixation of content (I have tried my own hand at developing a theory of LOT term content that assigns reference on the basis of a subject's actual history--see Rupert, forthcoming). Although their discussion is set in a much different context, Fodor and Lepore (1992, p. 157) give some sense of what is thought to be an obvious, general problem with an actual history-based causal theory of intentional content, the worry being that a person never has to engage in direct causal interactions with, say, tigers in order to token an LOT term that refers to them.

[5] Certain differences exist between the present project and Devitt's in that (a) Devitt is concerned primarily with natural language semantics and (b) Devitt is careful to separate questions about reference from more abstract questions about the nature of meaning itself (although note that Devitt believes truth and reference will ultimately emerge as the fundamental semantic properties, once the appropriate naturalistic methodology has been applied in semantics [Devitt 1994, p. 572]).

[6] See Churchland and Sejnowski 1992, p. 61, where the authors endorse similar methodology.

[7] While it seems that your average human can tell, with a fair degree of reliability, whether she is now having the same thought as she had on a previous occasion, this skill is of limited help in our development of a theory of LOT term individuation. This way of approaching LOT term individuation puts the cart before the horse: it individuates LOT terms according to their reference. If the naturalist is to locate content in the natural, i.e., nonsemantic, order of nature, then it seems that both terms of the 'refers to' relation will have to be nonsemantically specified. We should bear in mind that in cognitive science, mental representations have typically been thought of as a medium for the expression of content; intuitions regarding the content of a thought, while perhaps fairly reliable as indicators of thought content, cannot be taken as indicators of the presence of specific characters in the mental medium (though see note 3 above for discussion of the potential practical value of such intuitions).

[8] Localized portions of the brain seem to have specific jobs to do, and this suggests that the space to which we might look in search of a given LOT term could be of fewer dimensions than the one trillion or more suggested in the text. In any case, the number of free parameters will be quite large. Ignoring complications that result from the localized nature of cognitive processing should not materially affect the arguments that follow.

A contrasting complication arises if one takes seriously the idea that cognitive processing does not occur only in the brain (Rockwell 1994), and thus is much less localized than is often thought. The addition of 'noncranial' parameters would not, however, alter the arguments in the text.

[9] The preceding explication of dynamical systems theory draws heavily on Van Gelder and Port 1995, Norton 1995, and Giunti 1995. See these sources for a more complete explanation of the nature of dynamical systems and the mathematics used to analyze their behavior.

[10] I will use the abbreviation 'DS' only when talking about dynamical systems theory as it has been worked into a certain view of the mind, i.e., that the mind is most productively thought of as a dynamical system. When I mean instead to talk about dynamical systems theory in general, as an area of mathematics, I refer to it as such.

[11] Also see Horgan and Tienson's work (Horgan and Tienson 1994, 1996) for arguments to this effect. Horgan and Tienson's arguments differ from Van Gelder and Port's, though, in that Horgan and Tienson place most of their emphasis on difficulties arising from the frame problem and the related problem of the computational intractability of cognitive processing.

[12] For example, Jeffrey Elman says, "[R]epresentations are not abstract symbols but rather regions in a state space. Rules are not operations on symbols but rather embedded in the dynamics of the system, a dynamics which permits movement from certain regions to others while making other transitions difficult." (Elman 1995, p. 196) The mind does employ mental representations, and it does follow rules when processing these representations, Elman says, it's just that the representations and the rules "may be different from what we have conceived them to be." (ibid., p. 195)

[13] For examples of eliminativist claims coming from the DS camp, together with reasons why we should be suspicious of them, see Clark and Toribio 1994; also relevant are Clark 1991 and Fodor and Pylyshyn 1988, which offer criticisms of connectionist-inspired, eliminativist views that would also seem to hold against DS-inspired eliminativism. And for a recent attempt to develop some aspects of a noneliminativist, connectionist view of representation, see Churchland 1998.

[14] I do not mean to imply here that the infant's use of the LOT term 'object' is guided in *no way* by other representations; I merely suggest that the *content* of these other representations is not operative in fixing the extension of 'object' for the infant. The idea is that even if there is *some* degree of content-based interdependence in the determination of reference, some terms have to have their content fixed independently of the content-laden contribution of others; otherwise the entire process of content determination would, at least if one is a realist about content, not seem to ever be able to get off the ground. (This demand for the independent fixation of content for at least some LOT terms is similar to that made by Fodor and Lepore in their criticism of Paul Churchland's state-

space semantics [Fodor and Lepore 1996, pp. 152-5].) See Spelke 1990, 1991 and Bower 1989 for descriptions of experimental results that reveal the infant's early development of a strikingly rich object concept.

[15] I am not saying that Elman is confused about this point, only that this way of talking can be misleading.

[16] The reader may wonder whether we have some way of ruling out certain LOT terms as potential bearers of the content 'electron', other than saying that only attractors can correspond to LOT terms. One way to justify our judgements in this regard would be to provide a systematic, physiologically grounded distinction between physical structures that would be appropriate candidates for representing complex theoretical concepts and those that would be appropriate candidates for being mental representation of phonemes. As acknowledged in note 3 above, this kind of distinction is difficult to make at the physical level without employing some semantic hunches, although experimental, physiological data (such as those collected by questioning patients whose brains are directly stimulated while under local anaesthetic only) can help justify our general conclusions as to what types of physical structures would be appropriate to play what kinds of semantic roles in LOT. From a practical standpoint, however, we may be able to justify the judgement suggested in the text by observing that the subject in question fails to say sensible things about electrons and fails to exhibit other appropriately electron-directed behaviors.

[17] We should not ignore the possibility of innate contributions to the phase portrait. In developmental psychology, and cognitive science more generally (e.g., in psycholinguistics), nativism has many advocates (Karmiloff-Smith 1992 includes a largely approving review of much of the developmental data supporting nativism). One might think that the more cognitive development is governed by innate principles and constraints, the less plausible we should find an actual history-based theory of content for LOT terms, contrary to one of the main theses of the present work. (This type of worry has been raised with respect to Dretske's theory [Cummins 1991, pp. 105-6]). It is important to note, however, that as much as cognitive science has demonstrated in the way of innate cognitive biases, many, if not all, aspects of cognitive development are shaped by environmental interaction (This is one of Karmiloff-Smith's primary points in discussing the nativist literature), including the development of the wetware itself. The point here is that even if the newborn infant's phase space has some innate shape, the attractors that are to eventually count as

terms in LOT emerge only as the result of copious interaction with the environment. This point is especially clear when one considers the distinction between constraints on processing, the innate presence of which is supported by the nativist developmental work, and specific concepts explicitly represented at birth. Given the many forms an implicit rule or constraint can take (Cummins 1986), it seems that newborns could possess much in the way of innate biases, while having few, if any, innate, explicit mental representations (See Keil 1990 for a survey of the various types of constraints developmental researchers have in mind when they talk about innate constraints and biases).

[18] Bartsch 1996 offers a dynamics-based analysis of conceptual development amenable to the DS view outlined above. In Bartsch's model the stability of a concept depends on the emergence of structure in the cognitive system, which emergence of structure depends on the subject's direct experience of the world. Until a certain stability of structure emerges, the subject cannot be said to possess the concept in question, according to Bartsch. While Bartsch's analysis is largely in keeping with the DS view briefly described in the text, Bartsch distinguishes between concepts as structures present to consciousness and subcognitive structures that are not, and she does so in a way that only seems to give full representational status to structures of the former type (Bartsch 1996, pp. 430-31). In my discussion of LOT terms, however, I ignore this distinction, placing no special weight on conscious accessibility or lack thereof.

[19] Constructing a naturalistic semantics for LOT terms that is consistent with DS theory requires much work. In addition to locating a satisfactory content-fixing relation, we will have to include at least two following elements, relating specifically to DS theory: (a) a principle of individuation for LOT terms that separates those attractors that are to be identified with LOT terms from those which should not be, and (b) a principle of ancestry that tells us how attractors, as LOT terms, are to be identified over time, as the phase space in which they appear changes. Bartsch's model (Bartsch 1996) goes some distance toward characterizing the sort of stability in phase space we like to see before we attribute a concept to a given subject. However, a naturalistic semantic theory that assigns extensions on the basis of a subject's history also needs a way to identify currently stable attractors with their early ancestors in

order that the content of a current LOT term can be said to have been determined by past causal interactions involving its relevant ancestors.

[20] As noted above, the Crude Causal Theory has no real advocates, so far as I can tell. However, given the theoretical complexity of the causal history-based theories cited in note 4, it will be convenient to use the Crude Causal Theory to illustrate the need for a naturalistic theory of content to identify a suitable relation R.

[21] But see Andrew Pessin's worries about the mismatch between the causal powers of neural types and those of psychological states (Pessin 1995, p. 39 and passim).

[22] For classic statements of MR, see Putnam 1967 and Fodor 1974.

[23] Horgan and Tienson (1996, chapter 9, note 5) give a brief explanation of why it is empirically important to keep the 'multiples' of multiply realizable states fairly small. While they put their argument in terms of mental states, semantically interpreted, their points about the loss of empirical power would seem to hold when we limit the discussion to the multiple realizability of uninterpreted LOT terms.

[24] The text suggests a naturalism whose strictures require that all legitimate entities, properties, and relations be susceptible to characterization in nonintentional, nonsemantic terms; in contrast, some naturalistically minded philosophers reject such a restrictive view, in favor of a more liberal, ontologically promiscuous naturalism (Baker 1993, p. 94, Shapiro 1996, p. 541) that can take semantic properties as part of the basic furniture of the universe, in no need of reduction to or unification with other, nonsemantic theories of the universe. While I agree to some extent with Baker when she says that "Unity is merely desirable, not inevitable" (op. cit., p. 94), I think there are good reasons, which I cannot go into here, for a naturalist to find unity *highly*, not just "merely", desirable; thus, the naturalist should pursue as much theoretical unity as possible, though with an awareness that we may have to settle for the less desirable over what is decidedly more so. Even if the liberal naturalists were correct, it's not clear to what extent that would undermine my point in the text: even were we take semantic properties to be part of the basic

furniture of the universe, we would still somehow have to characterize the entities to which we wish to attribute those semantic properties; were we to individuate the entity that has semantic property *p* simply by referring to its possession of semantic property *p*, we would risk trivializing much of the explanatory work to be done by our semantic theory. In addition, a semantically laden criterion of identification goes against the apparent facts in the case of natural language, where, for the most part, we can individuate words according to their graphic and phonemic, i.e., nonsemantic, properties.

[25] The reader should bear in mind another sort of relativization to which the naturalist should be especially sensitive, one that I have hinted at already in the text. A naturalistic semantic theory should always be seen as part of an attempt to understand representation as a theoretical construct employed by a particular empirical theory (and folk theory is a possibility here) or set of connected theories (Cummins 1989, pp. 12-13). So not only should we relativize the discussion of a given theory of content to the theory(ies) used to individuate the relevant representing units (as argued in the text), our evaluation of a naturalistic semantic theory should also be constrained by the explanatory purposes to be served by our attribution of reference or intentional content: we should not criticize a naturalistic semantic theory for failing to do what it was not supposed to do in the first place--for example, for its failure to explain the source of each and every semantic intuition we might have about our thoughts.

[26] Sometimes Fodor claims only to be giving sufficient conditions for intentionality, in the attempt to show that Brentano is wrong to claim that intentionality is irreducible (Fodor 1990a, p. 96). But from the discussion of section II, the worry emerges that we are in no position to evaluate Fodor's alleged solution to Brentano's problem *unless* ADT applies to the best cases we know of intentional/representational systems, i.e., human beings. If ADT does not apply to the best (only?) cases of representational systems that we know of, what could be a naturalist's justification for claiming that asymmetric dependence is one way to fix intentional content, and thus provides a solution to Brentano's problem?

[27] Rarely does Fodor say much about the specific physical characteristics of these representations, beyond claiming that they can exhibit formal properties to which computational processes are sensitive (see, for example, Fodor

1994, Lecture 1). When Fodor does say more, it is not terribly enlightening. For example, in discussing an objection to ADT raised by Ned Block, Fodor says that for the purpose of talking about the tokening of LOT terms in counterfactual situations, we should understand the term 'cow' as a "phonological/orthographic sequence" (Fodor 1990a, p. 111). However, given that terms in LOT cannot be heard and are not written, this talk of phonology and orthography seems as if it would have to be metaphorical. My concern in the text is that Fodor does not offer a theory of LOT term individuation that might legitimate this metaphor.

The reader may be put off by my use of a quotation from Fodor 1990b, given that Fodor has openly renounced the views expressed there. While Fodor has given up on the teleological approach to LOT semantics described in 1990b, he has not given up his view of the physical constitution of LOT terms, so neatly summarized in the passage quoted in the text (cf. Fodor 1990a, p. 159).

[28] The empirical work I have in mind is the work on family resemblance and the prototype/stereotype structure of concepts. The early, ground-breaking work in this vein is summarized in Smith and Medin 1981. And though the past fifteen years have seen a mitigation of the more extreme claims of the prototype/stereotype-based view of concept structure (Armstrong, Gleitman, and Gleitman 1983, Smith 1989, Murphy and Medin 1985, Murphy 1988, Keil 1989), the basic point of this work seems to stand: humans rely largely on nondefinitional heuristics in applying concepts. For a recent discussion of the bearing of the nature of concepts on philosophical method, somewhat in keeping with the view expressed here, see Horgan and Tienson (1996, pp. 142-3). Putnam (1975) and Stich (1983) also provide provocative applications of a stereotype-based view of concepts to philosophical questions, about reference, in Putnam's case, and the nature of mental states, in Stich's (though by referring to these applications, I do not thereby endorse without qualification Putnam's or Stich's results). See Wittgenstein's discussions of concepts and meanings throughout his *Philosophical Investigations* (Wittgenstein 1953) for what would seem to be an intellectual ancestor of many contemporary views of the nature and structure of concepts.

By giving a limited endorsement of the stereotype/prototype view of concept structure, I do mean to imply that word meanings *are* stereotypes. (See Fodor 1981, Chpt. 10, and Fodor and Lepore 1992, Chpt. 6, for good reasons to doubt that stereotypes/prototypes can be word meanings, given one highly useful interpretation of 'meanings'.) Even if concepts with prototype structure help to attach some of our LOT or natural language terms to determinate

external meanings of the sort Fodor might favor (Fodor 1990a, Chpt. 6), the concepts themselves, as guides to the application of LOT or natural language terms, may lack such determinacy. (Note that I have kept separate concepts and LOT terms; this is because the concepts discussed in the psychological literature as stereotypes are typically complex in nature, whereas the LOT terms I have in mind are atomic terms. This leaves open the possibilities (1) that the typical concept discussed in the psychological literature is a collection of LOT terms--often called 'features'--and (2) that some concepts, the atomic ones, are such that each is identical to one atomic LOT term.)

[29] Cf. David Lewis's dismissal, in a different, though importantly similar, context, of the theoretical relevance of a hypothetical mad, unique Martian (Lewis 1980, p. 221).

[30] A referee for this journal brought to my attention the fact that both of these points are of particular concern in the present context.

[31] I have dealt only with the question of interpersonal comparison, not with intrapersonal comparisons. The latter should be handled analogously, though an additional complication arises when we wonder about the stability of attractors within a single subject over time: to fully explain how meaningful intrapersonal content comparisons can be made, we must decide on the appropriate way to characterize the structural continuities in a dynamical system whose phase space changes over time; see notes 18 and 19.

[32] There is a wealth of literature that addresses these questions. For a start, see the essays in Part I of Macdonald and Macdonald 1995; also valuable are Van Gelder 1990, Clark 1991, Clark and Toribio 1994, Horgan and Tienson 1996, Garson 1997, Ramsey 1997, and the works of DS theorists discussed in section IV above.

[33] There exist other mathematical methods that, when properly employed, may lead to the uncovering of syntactic structure or to the characterization of individual LOT terms within cognitive systems viewed as dynamical systems. The use of cluster analysis (Clark 1989, pp. 192-3) and tensor product encoding schemes (Smolensky 1991) are two such possibilities.

[34] Van Gelder and Port 1995, Clark 1995, and McClamrock 1995 argue that in one way or another, when we study human cognitive systems, knowledge of (what are thought of as) implementation details can have great bearing on our understanding of how the system works *at the cognitive level*.