The Role of Naturalness in Lewis's Theory of Meaning

Brian Weatherson

It is sometimes claimed (e.g., by Sider (2001a,b); Stalnaker (2004); Williams (2007); Weatherson (2003)) that David Lewis's theory of predicate meaning assigns a central role to naturalness.[1] Some of the people who claim this also say that the theory they attribute to Lewis is true. The authors I have mentioned aren't as explicit as each other about exactly which theory they are attributing to Lewis, but the rough intuitive idea is that the meaning of a predicate is the most natural property that is more-or-less consistent with the usage of the predicate. Call this kind of interpretation the 'orthodox' interpretation of Lewis.[2] Recently Wolfgang Schwarz (2009, 209ff) has argued that the orthodox interpretation is a misinterpretation, and actually naturalness plays a much smaller role in Lewis's theory of meaning than is standardly assumed.[3] Simplifying a lot, one key strand in Schwarz's interpretation is that naturalness plays no role in the theory of meaning in Lewis (1969, 1975), since Lewis hadn't formulated the concept yet, and Lewis didn't abandon that theory of meaning, since he never announced he was abandoning it, so naturalness doesn't play anything like the role orthodoxy assigns to it.

In this article I attempt to steer a middle ground between these two positions. I'm going to defend the following parcel of theses. These are all exegetical claims, but I'm also interested in defending the thesis that I ultimately attribute to Lewis, so getting clear on just what Lewis meant is of more than historical interest.

1. Naturalness matters to Lewis's (post-1983) theory of meaning only insofar as it matters to his theory of rationality, and the theory of rationality matters to the (pre- and post-1983) theory of meaning.
2. When we work through that theory of meaning, we see that the orthodox interpretation assigns to Lewis a theory that isn't his theory of meaning, but is by his lights a useful heuristic.

[1] Holton (2003) is more nuanced, but does tell a similar story in the context of discussing Lewis's account of (potential) semantic indeterminacy. Weatherson (2010) follows Holton in this respect.

[2] As some further evidence for how orthodox the 'orthodox' interpretation is, note that Williams (2007) is a prize winning essay published with two commentaries in the *Philosophical Review*. That paper takes the orthodox interpretation as its starting point, and neither of the commentaries (Bays (2007) and Hawthorne (2007)) criticises this starting point.

[3] Schwarz (2006) develops his criticism of orthodoxy in more detail, and in English, but it is as yet unpublished.

3. An even better heuristic than ‘meaning = use plus naturalness’ would be ‘meaning = predication plus naturalness’, but even this would be a fallible heuristic, not a theory.
4. When correctly interpreted, Lewis’s theory is invulnerable to the challenges put forward in Williams (2007).

I’m going to start by saying a little about the many roles naturalness plays in Lewis’s philosophy, and about what we might do if we were sceptical that one characteristic of properties could play all those roles. Then I’ll offer four arguments against the orthodox interpretation of Lewis, and in favour of the position described in the four numbered points above.

1 Naturalness in Lewis’s Philosophy

Most of the core elements of David Lewis’s philosophy were present, at least in outline, from his earliest work. The big exception is the theory of natural properties introduced in Lewis (1983a). As he says in that paper, he had previously believed that “set theory applied to possibilia is all the theory of properties that anyone could ever need” (Lewis, 1983a, 377n). Once he introduces this new concept of naturalness, Lewis puts it to all sorts of work. Much of that work is in metaphysics. For instance, in Lewis’s post-1983 metaphysics, naturalness plays the following roles, the first two of which are described in detail in Lewis (1983a), and the third of which appears in Lewis (1991, 1994a, 2001):

- Perfect duplication is defined in terms of sharing perfectly natural properties, and intrinsic properties are defined as those shared by perfect duplicates. (Though see the amendment to the 1983 theory in (Lewis, 1986, 61), and note that Langton and Lewis (1998) offers a theory of duplication and intrinsicness that uses naturalness in a slightly different way.)
- Laws are those generalisations that are strongest and simplest to state in a language where the predicates pick out perfectly natural properties.
- All the truths in the world supervene on the distribution of perfectly natural properties. As Lewis puts it, in a phrase borrowed from Bigelow (1988), “truth supervenes on being”.

Now it isn’t obvious that there is one division, into the natural and non-natural, that can play these roles, but clearly Lewis believes there is. Call the properties that are natural, in the sense needed for naturalness to play these roles, the M-natural properties.

Naturalness also plays some roles that are less distinctively metaphysical. For instance, naturalness plays a role in induction. It is rational to project greenness, and not to project gruesomeness. Call the properties that play this distinctively epistemological role the E-natural properties. Presumably, since he is interested in solving Goodman's riddle, he means that it is rational to make these projections without any prior information about the projectibility of greenness. And in any case we know, from Lewis (1994b), that he believed that there were some substantive foundational facts about what is and isn't rational. So it isn't absurd to attribute to Lewis a view on induction that's related to James Pryor's (2000) dogmatism about perception. The dogmatist view about induction I have in mind says that an agent is rational in inferring from the fact that observed Fs are G to the conclusion that unobserved Fs are Gs without any prior evidence that F is projectible provided (a) F is a natural property of the right kind, and (b) she doesn't have any evidence that F is not projectible. If such a theory is true, and again it does seem plausible to me, we can call the properties that play this distinctively epistemological role the E-natural properties.

Unless the M-natural properties *just are* the E-natural properties, then at least part of Lewis's theory of natural properties is not correct. But at least on first glance, it seems that these two notions are rather distinct. We can, as Goodman (1955) noted, rather easily make projections about colour. But colours are not particularly M-natural; they don't make for much objective similarity, they don't play a central role in the theory of laws and so on. On the other hand, some properties that are natural might not be easily projectible. The fact that electronhood is very natural doesn't in itself make it easy to project. Something nearby to this is true: if we *know* that electronhood is natural, then we can easily project it. If one knows electronhood is highly natural, one can go from evidence about one electron's mass to conclusions about another's mass. But I don't see how this would be rational in the absence of such knowledge. So I conclude, a little tentatively, that M-naturalness and E-naturalness are distinct. I'll make that conclusion a little less tentaive later in this section.

That leads to a difficult interpretative question. Assume that Lewis was convinced that he should distinguish E-naturalness from M-naturalness. This could be because he was convinced that the arguments of the previous two paragraphs were correct. Or, perhaps in a more nearby possible world, it could be because he was convinced that enough people believed the arguments of the previous section that he shouldn't presuppose the identity of E-naturalness and M-naturalness. (Lewis preferred, where possible, to make clear which of his views could be held without presupposing his other views.) Then, if orthodoxy was right and Lewis wanted to include naturalness in his theory of meaning, would he say it was E-naturalness or M-naturalness that was to be included?[4] Or, if he were convinced that there was not a unified theory of naturalness, would he conclude that we also needed a separate semantic concept, call it S-naturalness, to play the role of naturalness in his theory of meaning.

I think there are both textual reasons and theoretical reasons for saying that E-naturalness is what does the work in Lewis's theory of meaning. The textual reason is that in Lewis (1992), he describes Goodman's puzzle and the Kripkenstein puzzle as the same puzzle. In the orthodox story at least, naturalness comes in to solve the Kripkenstein puzzle. And, of course, it is E-naturalness that solves Goodman's puzzle. So it must be E-naturalness that solves the Kripkenstein puzzle.

The theoretical reasons are also interesting. To see them, let's start with three Lewisian themes.

- Facts about linguistic meaning are to be explained in terms of facts about minds. In particular, to speak a language $\mathscr{L}$ is to have a convention of being truthful and trusting in $\mathscr{L}$ (Lewis, 1969, 1975). And to have such a convention is a matter of having certain beliefs and desires. So mental content is considerably prior to linguistic content in a Lewisian theory. Moreover, Lewis's theory of linguistic content is, in the first instance, a theory of *sentence* meaning, not a theory of *word* meaning. So any attribution of a theory of word meaning to Lewis involves some attributions of a theory that isn't made explicit in Lewis.[5]
- The principle of charity plays a central role in Lewis's theory of mental content Lewis (1974, 1994b). To a first approximation, a creature believes that p iff the best interpretation of the creature's behavioural dispositions includes the attribution of the belief that p to the creature. And, ceteris paribus, it is better to interpret a creature so that it is more rather than less rational. It will be pretty important for what follows that Lewis adopts a principle of charity that highlights *rationality*, not *truth*. It is also important to Lewis that we don't just interpret the individual creature, but

[4] Williams (2007) connects Lewis's preference for natural interpretations to his preference for simple laws, where simplicity is measured in terms of ease of statability in a perfectly natural language. In the terms of this paper, this suggests that it is M-naturalness that matters to meaning, not E-naturalness. I'm going to argue that this is a mistake.

[5] These points are stressed by Wolfgang Schwarz (2006, 2009). He also notes that in "Putnam's Paradox" Lewis explicitly sets these parts of his theory aside so he can discuss Putnam's arguments on grounds most favourable to Putnam. As Schwarz says, this should make us suspicious of the central role "Putnam's Paradox" plays in defences of the orthodox interpretation. We will return to this point in the section on textual evidence for and against orthodoxy.

creatures of a kind (Lewis, 1980). I'm not going to focus on the social externalist features of Lewis's theory of mental states, but I think they assist the broader story I want to tell.
- Lewis's theory of mental content has it that mental contents are (what most of us would call) properties, not (what most of us would call) propositions (Lewis, 1979). So a theory of E-natural properties can easily play a role in the theory of content. To say that a property is more E-natural is just to say that an agent needs less evidence to believe it than she needs to believe equally strong propositions that are more E-natural. If you don't like the Lewisian theory of mental content (and it is extremely controversial) then we need a theory of E-natural propositions, not just (or perhaps instead of) a theory of E-natural properties. There is no easy map from E-naturalness of properties to E-naturalness of propositions. That's because it doesn't take much evidence to rationally believe propositions like ⟨⟨All emeroses are gred⟩⟩, which has very unnatural constituents.

Now let's see why we might end up with naturalness in the theory of meaning. An agent has certain dispositions. For instance, after seeing a bunch of green emeralds, and no non-green emeralds, in a large and diverse range of environments, she has a disposition to say "All emeralds are green". In virtue of what is she speaking a language in which "green" means green, and not grue? (Note that when *I* use "grue", I mean a property that only differs from greenness among objects which it is easy to tell that neither our agent, nor any of her interlocutors, could possibly be acquainted with at the time she makes the utterance in question.)

Let's say that $\mathscr{L}_1$ is English, i.e., a language in which "green" means green, and $\mathscr{L}_2$ a language which is similar to $\mathscr{L}_1$ except that "green" means grue. Our question is, what makes it the case that the agent is speaking $\mathscr{L}_1$ and not $\mathscr{L}_2$? That is, what makes it the case that the agent has adopted the convention of being truthful and trusting in $\mathscr{L}_1$, and not the convention or being truthful and trusting in $\mathscr{L}_2$? On a theory that adopts the bulleted points listed above, it must be E-naturalness. Let's look at why this must be so.

We assumed that the agent has seen a lot of emeralds which are both green and grue. To a first approximation, it is more charitable to attribute to the agent the belief that all emeralds are green than the belief that all emeralds are grue because greenness is more natural than gruesomeness. As Lewis says, "The principles of charity will impute a bias towards believing things are green rather than grue" (1983a, 375). And for Lewis, charity requires imputing more reasonable interpretations. But why is it more charitable to attribute beliefs about greenness to beliefs about grueness? I think it is because we need more evidence to rationally form a belief that some class of things are all grue than we need to form a belief that everything in that class is green. The agent has, we might assume, sufficient evidence to rationally believe that all emeralds are green, but not sufficient evidence to believe that all emeralds are grue.

So the first two Lewisian themes notes above, the reduction of linguistic meaning to mental content, and the centrality of a rationality-based principle of charity, push us towards thinking that E-naturalness is closely connected to mental content and hence to linguistic meaning. But the argument we offered was a bit quick, because we forgot the third theme: beliefs are relations to properties, not propositions. It will turn out that taking this theme seriously will actually *strengthen* the case that E-naturalness plays a central role in Lewis's theory of meaning, and strengthen the case that Lewis should have distinguished E-naturalness from M-naturalness.

If a certain body of evidence makes it possible for the agent to rationally believe that all emeralds are green, but not for her to believe that all emeralds are grue, then that must be because the first of the following properties is more natural than the second:

- Being in a world where all emeralds are green
- Being in a world where all emeralds are grue

Indeed, the first of these is intuitively *much more* E-natural than the second. But it isn't considerably more M-natural than the second. Indeed it isn't clear that it is more M-natural *at all*. Neither of these properties makes for any significant kind of objective similarity; the first property is, after all, shared by everything under the sun, and much more besides. Neither of these properties will play any kind of role in a law, and so on. So if the first is much more E-natural than the second, and it must be for Lewis's theory to Goodman's riddle to work, E-naturalness and M-naturalness must come apart. And if we are to explain how "green" means green and not grue in Lewis's characteristic way, i.e., via reducing linguistic meaning to mental content, and determining mental content by the principle of charity, it must also be that it is this difference in E-naturalness, not any marginal difference in M-naturalness, that is relevant to meaning.

So this is an argument both that M-naturalness and E-naturalness come apart and, since we want the theory of meaning to tell us that "green" means green and not grue, that it is E-naturalness that matters for Lewis's account of content.

The bias (Lewis, 1983a, 375) talks about is a bias towards the E-natural, not the M-natural, when these come apart.[6]

Given this, if the orthodox interpretation of Lewis is correct, it must be that the meaning of a predicate is the most *E-natural* property that is (more-or-less) consistent with the usage of the word. We'll now turn to four problems with this theory of meaning, starting with the textual evidence for and against it.

2 Textual Evidence

There is some *prima facie* textual evidence for the orthodox interpretation. But looking more careful at the context of these texts not just undermines the support the text gives to the orthodox interpretation, but actually tells against it. (This part of the paper is indebted even more than the rest to Wolfgang Schwarz's work, and could be easily skipped by those familiar with that work.)

I'll focus on the last seven pages of "New Work for a Theory of Universals". This is the part of "New Work" that uses the notion of naturalness, as introduced in the paper, to respond to Putnam's model-theoretic arguments for massive indeterminacy of meaning. Lewis actually responds to Putnam twice over. First, he responds to Putnam directly, by showing how adding naturalness to a use-based theory of sentence meaning avoids the 'just more theory' objection that's central to Putnam's argument. And when Lewis describes this direct response, he says things that sound a lot like the orthodox interpretation.

> I would instead propose that the saving constraint concerns the referent - not the referrer, and not the causal channels between the two. It takes two to make a reference, and we will not find the constraint if we look for it always on the wrong side of the relationship. Reference consists in part of what we do in language or thought when we refer, but in part it consists in eligibility of the referent. And this eligibility to be referred to is a matter of natural properties. (Lewis, 1983a, 371)

But after this direct response is finished, Lewis notes that he has conceded quite a lot to Putnam in making the response.

[6] Note that I am *not* claiming that Lewis himself distinguished E-naturalness from M-naturalness. The failure to do this led to certain complications, such as those discussed on page 228 of Lewis (1984). I am just saying that had he decided to distinguish these, he would have not seen any further need to add another concept of naturalness to explain the role of naturalness in meaning, and he would have said it was E-naturalness, not M-naturalness, which mattered for the theory of meaning.

> You might well protest that Putnam's problem is misconceived, wherefore no need has been demonstrated for resources to solve it. ... Where are the communicative intentions and the mutual expectations that seem to have so much to do with what we mean? In fact, where is thought? ... I think the point is well taken, but I think it doesn't matter. If the problem of intentionality is rightly posed there will still be a threat of radical indeterminacy, there will still be a need for saving constraints, there will still be a remedy analogous to Merrill's suggested answer to Putnam, and there will still be a need for natural properties. (Lewis, 1983a, 373)

I noted earlier that Schwarz makes much of a similar passage in "Putnam's Paradox", and I think he is right to do so. Here's a crucial quote from that paper.

> I shall acquiesce in Putnam's linguistic turn: I shall discuss the semantic interpretation of language rather than the assignment of content to attitudes, thus ignoring the possibility that the latter settles the former. It would be better, I think, to start with the attitudes and go on to language. But I think that would relocate, rather than avoid, the problem; wherefore I may as well discuss it on Putnam's own terms. (Lewis, 1984, 222)

That passage ends with a footnote where he says the final section of "New Work" contains a version of how the 'relocated' problem would be solved. So let's turn back to that. The following long portmanteau quote from pages 373 to 375 captures, I think, the heart of my interpretation.

> The problem of assigning content to functionally characterised states is to be solved by means of constraining principles. Foremost among these are principles of fit. ... A state typically caused by round things before the eyes is a good candidate for interpretation as the visual experience of confronting something round; and its typical impact on the states interpreted as systems of belief ought to be interpreted as the exogenous addition of a belief that one is confronting something round, with whatever adjustment that addition calls for. ... Call two worlds equivalent iff they are alike in respect of the subject's evidence and behaviour, and note that any decent world is equivalent inter alia to horrendously counterinductive worlds and to worlds where everything unobserved by the subject is horrendously nasty. ... We can interchange equivalent worlds ad lib and preserve fit. So,

> given any fitting and reasonable interpretation, we can transform it into an equally fitting perverse interpretation by swapping equivalent worlds around . . . If we rely on principles of fit to do the whole job, we can expect radical indeterminacy of interpretation. We need further constraints, of the sort called principles of (sophisticated) charity, or of 'humanity'. [A footnote here refers to "Radical Interpretation".] Such principles call for interpretations according to which the subject has attitudes that we would deem reasonable for one who has lived the life that he has lived. (Unlike principles of crude charity, they call for imputations of error if he has lived under deceptive conditions.) These principles select among conflicting interpretations that equally well conform to the principles of fit. They impose *apriori* – albeit defeasible - presumptions about what sorts of things are apt to be believed and desired . . . **It is here that we need natural properties.** The principles of charity will impute a bias toward believing that things are green rather than grue . . . In short, they will impute eligible content . . . They will impute other things as well, but it is the imputed eligibility that matters to us at present. (Lewis, 1983a, 373-5, my emphasis)

I think that does a reasonably clear job of supporting the interpretation I set out in the introduction over the orthodox interpretation. Naturalness matters to linguistic meaning all right. But the chain of influence is very long and indirect. Naturalness constrains what is reasonable, reasonableness constrains charitable interpretations, charitable interpretations constrain mental content, and mental content constrains linguistic content. Without naturalness at the first step, we get excessive indeterminacy of content. With it, the Putnamian problems are solved. But there's no reason to think naturalness has any more direct role to play at any level in the theory of linguistic content.

In short, Lewis changed what he thought about rationality when he adopted the theory of natural properties. Since rationality was a part of his theory of mental content, and mental content determines linguistic content, this change had downstream consequences for what he said about linguistic content. But there wasn't any other way his theory of linguistic content changed, nor, contra orthodoxy, any direct link between naturalness and predicate meaning.

Moreover, when we look at the closest thing to a worked example in Lewis (1984), we don't get any motivation for the orthodox interpretation. Here's the example he uses, which concerns mental content. Let f be any mapping from worlds to worlds such that the agent has the same evidence and behaviour in w and $f(w)$. Extend f to a mapping from sets of worlds to sets of worlds in the following (standard) way: $f(S) = \{f(w) : w \in S\}$. Then the agent's behaviour will be rationalised by her evidence just as much if she has credence function C and value function V, as if she has credence function C' and value function V', where $C'(f(S)) = C(S)$, and $V'(f(S)) = V(S)$. To relate this back to the familiar Goodmanian puzzle, let f map any world where all emeralds are green to nearest world where all emeralds are grue, and vice versa, and map any other world to itself. Then the above argument will say that the agent's behaviour is rationalised by her evidence just as much as if her credences are C as if they are C'. That is, her behaviour is rationalised by her evidence just as much if she gives very high credence to all emeralds being green as to all emeralds being grue. So understanding charity merely as rationalizing behaviour leaves us without a way to say that the agent believes unobserved emeralds are green and not grue.

Lewis's solution is to say that charity requires more than that. In particular, it requires that we assign natural rather than unnatural beliefs to agents where that is possible. I've argued above that this makes perfect sense if we understand the notion of naturalness to be E-naturalness. The crucial thing to note here is that this all happens a long time before we can set out the way that a sentence is used, since the way a sentence is used on Lewis's theory of linguistic content includes the beliefs that are formed on hearing it. So the discussion in "New Work" suggests that naturalness matters for content, but not in a way that can be easily factorised out. And that's exactly what I think is the best way to understand Lewis's theory.

3 An Argument for the Orthodox Interpretation

Perhaps there is a more indirect way to motivate the orthodox interpretation of Lewis. Once we've distinguished M-naturalness from E-naturalness, the orthodox interpretation attributes to Lewis a theory that is quite attractive as a theory of semantic determinacy and indeterminacy. Call that theory the **U&N Theory**, short for the **U**se plus **N**aturalness theory of meaning. Since Lewis was clearly looking for such a theory when he discussed naturalness in the context of his theory of content, it is reasonably charitable to attribute the **U&N Theory** to him, as the orthodox interpretation does.

My response to this will be in three parts. First, I'll argue in this section that my rival interpretation attributes to Lewis a theory of semantic determinacy and indeterminacy that does just as well at capturing the facts Lewis wanted a theory to capture, so there's no charity based reason to attribute the **U&N Theory** to him (And, as we saw in the previous section, there's no direct textual reason to attribute it to him either.) Second, the **U&N Theory** is subject to the criticisms in Williams (2007), while the theory I attribute to Lewis is not. Third, the **U** part of the **U&N Theory** is hopelessly vague; it isn't clear how to say what 'use' is on a Lewisian theory that makes it suitable to add to naturalness to deliver meanings. Either use is so thick that naturalness is unneeded, or it is so thin that naturalness won't be sufficient to set meaning.So actually it isn't particularly charitable to attribute this theory to him.

Still, let's start with the attractions of the **U&N Theory**. On the one hand, agents are inclined to say "All emeralds are green" both in situations where they've seen a lot of green emeralds (and no non-green ones) and in situations where they've seen a lot of grue emeralds (and no non-grue ones). That's because, of course, those are exactly the same situations. So at first glance, it doesn't look like the way in which "green" is used will determine whether it means green or grue. On the other hand, once we add a requirement that terms have a relatively natural meaning, we do get this to fall out as a result. Moreover, given the M-naturalness/E-naturalness distinction, we can even see how this falls out of a recognisably Lewisian approach to meaning.

Consider again our agent who says "All emeralds are green" after seeing a lot of emeralds that are both green and grue. And remember that for her to speak a language, she must typically conform to conventions of truthfulness and trust in that language. Now if the agent was speaking $\mathscr{L}_2$, she would have to think that she's doing an OK job of being truthful in $\mathscr{L}_2$ by saying "All emeralds are green". But that would be crazy. Why should she think that all emeralds are grue given her evidence base? To attribute to her that belief would be to gratuitously attribute irrational beliefs to her. And on Lewis's picture, gratuitous attributions of irrationality are false. So the agent doesn't have that belief. So she's not speaking $\mathscr{L}_2$.

Things are even clearer from the perspective of hearers. A hearer of "All emeralds are green" would be completely crazy to come to believe that all emeralds are grue. The hearer knows, after all, that the speaker has no acquaintance with the emeralds that would have to be blue for all emeralds to be grue. So the hearer knows that this utterance could not be sufficient evidence to believe that all emeralds are grue. Yet if she speaks $\mathscr{L}_2$, she is disposed to believe that all emeralds are grue on hearing "All emeralds are green". She isn't irrational, or at least we shouldn't assign irrationality to her so quickly, so she doesn't speak $\mathscr{L}_2$.

So it looks like in this one case at least, we have a case where use plus naturalness gives us the right theory. Agents are disposed to use "green" to describe emeralds that are green/grue. But the fact that greenness is more natural than gruesomeness makes it more appropriate to attribute to them a convention according to which "All emeralds are green" means that all emeralds are green and not that all emeralds are grue.

But more carefully, what we should say is that the **U&N Theory** gives us the right result in this case. It doesn't follow that it will work in all cases, or anything like it. And it doesn't follow that it works for the right reasons. As we'll see, neither of those claims are true. In fact, just re-reading the last three paragraphs should undermine the second claim. Because we just saw a derivation that the agents are not speaking $\mathscr{L}_2$, that didn't even appeal to the **U&N Theory**. Rather, that derivation simply used the theory of meaning in *Convention* and the theory of mental content in "Radical Interpretation". It's true that the latter theory assigns a special role to rationality, and the theory of rationality we used has, among other things, a role for natural properties, but that is very different to the idea that naturalness feeds directly into the theory of meaning in the way the orthodox interpretation says. As I said at the start, I think the best interpretation of Lewis is that he changed his theory of *rationality* in 1983, but that's the only change to his theory of *meaning*.

Put another way, these reflections on "green" and "grue" are consistent with the view that the **U&N Theory** is a false *theory*, but a useful *heuristic*. It's a useful heuristic because it agrees with the true Lewisian theory in core cases, and is much easier to apply. That's exactly what I think the **U&N Theory** is, both as a matter of fact, and as a matter of Lewis interpretation.

4 Indeterminacy and Radically Deviant Interpretations

If the **U&N Theory** is a heuristic not a theory, we should expect that it will break down in extreme cases. That's exactly what we see in the cases discussed in Williams (2007). Those cases highlight the fact that a Lewisian theorist needs to be careful that we don't end up concluding that normal people, such as the agent in our example who says "All emeralds are green", speak $\mathscr{L}_3$. $\mathscr{L}_3$ is a language in which all sentences express claims about a particular mathematical model (essentially a Henkin model of the sentence the agent accepts), and it is set up in such a way that ordinary English sentences come out true, and about very natural parts of the model. On the **U&N Theory**, it could easily turn out that ordinary speakers are speaking $\mathscr{L}_3$, since the assigned meanings are so natural. We can see this isn't a consequence of *Lewis's* theory by working through the case from first principles. I have two arguments here, the first of them relying on some slightly contentious claims about the epistemology of mathematics, the second less contentious.

Assume, for reductio, that ordinary speakers are speaking $\mathscr{L}_3$. So, for instance, when O'Leary says "The beer is in the fridge", what he says is that a certain complicated mathematical model has a certain property. (And indeed it has that property.) Now this won't be a particularly rational thing for O'Leary to say unless he knows more mathematics than ordinary folks like him ordinarily do. So if O'Leary has adopted a convention of truthfulness and trust in $\mathscr{L}_3$, then uttering "The beer is in the fridge" would be irrational, even if he is standing in front of the open fridge, looking at the beer. That's a gratuitous assignment of irrationality, and gratuitous assignments of irrationality are false, so O'Leary doesn't speak $\mathscr{L}_3$.

Perhaps that is too quick. After all, the mathematical claim that $\mathscr{L}_3$ associates with "The beer is in the fridge" is a necessary truth. And Lewis's theory of content is intentional, not hyper-intentional. So O'Leary does know it is true. (And when he is standing in front of the fridge, there's even a sense that he knows that "The beer is in the fridge" expresses a truth, if $\mathscr{L}_3$ is really his language.) I think that's probably not the right sense of "rational", and I'm not altogether sure how much hostility to hyper-intensionalism we should attribute to Lewis. But so as to avoid these questions, it's easier to consider a different argument that focusses attention on O'Leary's audience.

When O'Leary says "The beer is in the fridge", Daniels hears him, and then walks to the fridge. Why does Daniels make such a walk? Well, he wants beer, and believes it is in the fridge. That looks like a nice rational explanation. But why does he believe the beer is in the fridge? I say it's because he's (rationally) adopted a convention of truthfulness and trust in $\mathscr{L}_1$, and so he rationally comes to believe the beer is in the fridge when O'Leary says "The beer is in the fridge". On the assumption that O'Leary and Daniels speak $\mathscr{L}_3$, none of this story goes through. But we must have some rational explanation of why O'Leary's statement makes Daniels walk to the fridge. So O'Leary and Daniels must not be speaking $\mathscr{L}_3$.

Michael Morreau pointed out (when I presented this talk at CSMN) that the preceding argument may be too quick. Perhaps there is a way of rationalising Daniels's actions upon hearing O'Leary's words consistent with the idea that they both speak $\mathscr{L}_3$. Perhaps, for instance, Daniels's walking to the fridge constitutes saying something in a complicated sign language, and that thing is the rational reply to what O'Leary said. If this kind of response works, and I have no reason to think it won't, the solution is to increase the costs to Daniels of performing such a reply. For instance, not too long ago I heard Mayor Bloomberg say "Lower Manhattan is being evacuated because of the impending hurricane", and I (and my family) packed up and evacuated from Lower Manhattan. Even if one could find an interpretation of our actions in evacuating that made them constitute the assertion of a sensible reply to Bloomberg's mathematical assertion in $\mathscr{L}_3$, it would be irrational to think I made such an assertion. Evacuating ahead of a storm with an infant is not fun - if it was that hard to make mathematical assertions, I wouldn't make them! And I certainly wouldn't make them in reply to someone who wouldn't even see my gestures. So I think at least some of the actions that are rationalised by testimony, interpreted as sentences of $\mathscr{L}_1$, are not rationalised by testimony, interpreted as $\mathscr{L}_3$. By the kind of appeal to the principle of charity we have used a lot already, that means that L_3 is not the language most people speak.

The central point here is that when we are ruling out particularly deviant interpretations of some speakers, we have to make heavy use of the requirement that the interpretation of their shared language rationalises what they do. In part that means it must rationalise why they utter the strings that they do in fact utter. And when we're considering this, we should remember the role of E-naturalness in a theory of rationality. But it also means that it must rationalise why people respond to various strings with non-linguistic actions, such as walking to the fridge, or evacuating Lower Manhattan. Naturalness has less of a role to play here, but the Lewisian theory still gets the right answers provided we apply it carefully. Since the Lewisian theory gets the right answers, and the **U&N Theory** gets the wrong answers, it follows that the **U&N Theory** isn't Lewis's theory, and so orthodoxy is wrong.

5 What is the Use of a Predicate?

We concluded the last section with an argument that Lewis isn't vulnerable to the claim that his theory assigns complicated mathematical claims as the meanings of ordinary English sentences. That interpretation, we argued, is inconsistent with the way those sentences are used. In particular, it is inconsistent with the way that *hearers* use sentences to guide their actions.

So far so good, we might think. But notice how much has been packed into the notion of use to get us this far. In identifying the use O'Leary makes of "The beer is in the fridge", we have to say a lot about O'Leary's beliefs and desires. And in identifying the use Daniels makes of it, we *primarily* talk about the sentence's effects on Daniels's beliefs and desires. That is, just saying how the sentence is used requires saying a lot about mental states of speakers. And that will often require appealing to constitutive rationality; we say that Daniels's beliefs about the fridge changed because we need to rationalise his fridge-directed behaviour.

And this should all make us suspicious about the prospects for identifying meaning (in a Lewisian theory) with use plus naturalness. The argument above that naturalness mattered to meaning relied on the idea that E-naturalness matters because it affects which states are rational, and hence which states are actualised. A belief that all emeralds are grue is unnatural (i.e., un-E-natural), so it is hard to hold. And since it is hard to hold, it is hard to think one is conforming to a convention of truthfulness in a language if one utters sentences that mean, in that language, that all emeralds are grue. That's why it is wrong, *ceteris paribus*, to interpret people as speaking about grueness.

But now consider what happened when we were talking about Daniels and O'Leary. Even to say how they were using the sentence "The beer is in the fridge", we had to say what they believed before and after the sentence was uttered. In other words, their mental states were constitutive of the way the sentence was used. Now add in the extra premise, argued for above, that naturalness matters to Lewis's theory of linguistic content because, and only because, it matters to his theory of mental content. (And it only matters to mental content because it matters to the principle of charity that Lewis uses.) If mental states, and their changes, are part of how the sentences are used, it will be rather misleading to say that meaning is determined by use plus naturalness. A better thing to say is that meaning is determined by use, and that some key parts of use, i.e., mental states of speakers and hearers, are determined in part by naturalness.

So I'm sceptical of the **U&N Theory**. We can put the argument of the last few paragraphs as a dilemma. There are richer and thinner ways of identifying the use to which a sentence is put. A thin way might, for instance, just focus on the observable state of the part of the physical world in which the sentence is uttered. A rich way might include include, inter alia, the use that is made of the sentence in the management of belief and the generation of rational action. If we adopt the thin way of thinking about use, then adding naturalness won't be enough to say what makes it the case that O'Leary and Daniels are speaking $\mathscr{L}_1$ rather than $\mathscr{L}_3$. If we adopt the rich way of thinking about use, then the role that naturalness plays in the theory of meaning has been incorporated into the metaphysics of use. Neither way makes the **U&N Theory** true while assigning naturalness an independent role. This dilemma isn't just an argument that we shouldn't attribute the **U&N Theory** to Lewis; it is an argument against anyone adopting that theory.

6 From Theory to Applied Semantics

So far we've argued that Lewis's semantic theory did not look a lot like the orthodox interpretation. It's true that he thought the way a sentence was used was of primary importance in determining its meaning. And it's true that he thought naturalness mattered to meaning. But that wasn't because naturalness came in to resolve the indeterminacy left in a use-based theory of meaning. Rather, it was because naturalness (more precisely, E-naturalness) was in a part of the theory of mental content, and specifying the mental states of speakers and hearers is part of specifying how the sentence is used.

But note that these considerations apply primarily to investigations at a very high level of generality, such as when we're trying to solve the problems described in "Radical Interpretation". They don't apply to investigations into applied semantics. Let's say we are trying to figure out what O'Leary and Daniels mean by "green". And assume that we are taking for granted that they are speaking a language which is, in most respects, like English. This is hardly unusual in ordinary work in applied semantics. If we are writing a paper on the semantics of colour terms, a paper like, say, "Naming the Colours", we don't concern ourselves with the possibility that every sentence in the language refers to some complicated mathematical claim or other.

Now given those assumptions, we can identify a moderately thin notion of use. We know that O'Leary uses "green" to describe things that are, by appearance, both green and grue. We also know that when O'Leary makes such a description, Daniels expects the object will be both green and grue. So focus on a notion of use such that the *use* of a predicate just is a function of which objects speakers will typically apply the predicate to, and which properties hearers take those objects to have once they hear the predication. If we wanted to be more precise, we could call this notion of 'use' simply *predication*. When we are doing applied semantics, especially when we are trying to figure out the meaning of predicates, we typically know which objects a speaker is disposed to predicate a predicate of, and that's the salient feature of use. (This is why I said the most accurate heuristic would be meaning is predication plus naturalness; predication is the bit of use we care about in this context.)

This identification of use wouldn't make any sense if we were engaged in theorising at a much more abstract level. If we are doing radical interpretation, then we have to take non-semantic inputs, and solve simultaneously for the values of the subject term and the predicate term in a (simple) sentence. But when we are just doing applied semantics, and working just on the meaning of a term like "green" in a well-functioning language, we can presuppose facts about the denotation of the subject term in sentences like *S is green*, and presuppose facts about what is the subject and what is the predicate in that sentence, and then we can look at which properties hearers come to associate with that very object on hearing that sentence.

Now that we have a notion of use that's distinct from naturalness, we can ask whether it is plausible that predicate meaning is use (in that sense) plus naturalness. And, quite plausibly, the answer is yes. The arguments in Sider (2001a) and Weatherson (2003) in favour of this theory look like, at the very least, good arguments that the theory does the right job in resolving Kripkensteinian problems. The theory is immune to objections based on radical re-interpretations of the language, as in Williams (2007), because those will be inconsistent with the use so defined. And the theory fits nicely into Lewis's broader theory of meaning, i.e., his metasemantics, which is in turn well motivated. So I think there are good reasons to hold that when we're doing applied semantics, the **U&N Theory** delivers the right verdicts, and delivers them for Lewisian reasons. That's the heart of what's true about the **U&N Theory**, even if it isn't a fully general theory of meaning.

7 Conclusion

I've concluded that the orthodox interpretation is, at best, a misleading description of Lewis's approach to meaning, although it does describe fairly accurately the consequences of his meaning theory for some puzzles in applied semantics. I'm going to conclude by making two observations about how the E-naturalness/M-naturalness distinction affects our view of Lewis's overall theory.

First, there is a quick argument that everyone needs some kind of theory of E-naturalness or other. If we didn't think that "green" denoted something more E-natural than all the possible denotations of "grue", "gred", "grurple" and so on, we would think that seeing a whole bunch of emeralds which are green and grue and gred and grurple and so on would give us equal reason to believe the next emerald we see will be green and grue and gred and grurple and so on. But if we now take these neologisms to denote properties like *green if observed by me, purple if not*, that will mean that all our evidence about emeralds gives us equal reason to believe that the next emerald we see will be green, blue, red, purple and so on. In other words, we can never get inductive evidence about the colour of unobserved emeralds. And since this little argument generalises pretty quickly, that means we never get inductive evidence for anything. So the alternative to believing in E-naturalness is a strong kind of inductive scepticism. And that's absurd, so we have to believe in E-naturalness.

Second, though, E-naturalness could be very different to M-naturalness. We don't have to think, as Lewis says in Lewis (1986), that E-naturalness is invariant across worlds, or knowable *a priori*. We could hold that the E-naturalness of a property is related to whether it is part of a homeostatic property cluster (Boyd, 1988). Or we could follow conservative approaches to epistemology, and say that how E-naturalness is related to how we humans do in fact project properties (Goodman, 1955; Harman, 1986). If we do think of E-naturalness this way, and incorporate it into our theory of meaning, we might end up linking meaning closely to cognitively primitive generalisations, as in Leslie (2008). This use of naturalness, and the resulting theory, would be a long way from the impression Lewis gives of his theory. But I think it is a sensible outcome of taking Lewis's views about the nature of mind and language, and mixing them with a different epistemology.

References

Bays, Timothy. 2007. "The Problem with Charlie: Some Remarks on Putnam, Lewis and Williams." *Philosophical Review* 116:401–425, doi:10.1215/00318108-2007-003.

Bigelow, John. 1988. *The Reality of Numbers: A Physicalist's Philosophy of Mathematics*. Oxford: Oxford.

Boyd, Richard. 1988. "How to Be a Moral Realist." In Geoffrey Sayre-McCord (ed.), *Essays in Moral Realism*, 181–228. Ithaca: Cornell University Press.

Goodman, Nelson. 1955. *Fact, Fiction and Forecast,*. Cambridge: Harvard University Press.

Harman, Gilbert. 1986. *Change in View*. Cambridge, MA: Bradford.

Hawthorne, John. 2007. "Craziness and Metasemantics." *Philosophical Review* 116:427–440, doi:10.1215/00318108-2007-004.

Holton, Richard. 2003. "David Lewis's Philosophy of Language." *Mind and Language* 18:286–295, doi:10.1111/1468-0017.00228.

Langton, Rae and Lewis, David. 1998. "Defining 'Intrinsic'." *Philosophy and Phenomenological Research* 58:333–345. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 116-132.

Leslie, Sarah-Jane. 2008. "Generics: Cognition and Acquisition." *Philosophical Review* 117:1–47, doi:10.1215/00318108-2007-023.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

—. 1974. "Radical Interpretation." *Synthese* 27:331–344. Reprinted in *Philosophical Papers*, Volume I, pp. 108-118.

—. 1975. "Languages and Language." In *Minnesota Studies in the Philosophy of Science*, volume 7, 3–35. Minneapolis: University of Minnesota Press. Reprinted in *Philosophical Papers*, Volume I, pp. 163-188.

—. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review* 88:513–543. Reprinted in *Philosophical Papers*, Volume I, pp. 133-156.

—. 1980. "Mad Pain and Martian Pain." In Ned Block (ed.), *Readings in the Philosophy of Psychology*, volume I, 216–232. Cambridge: Harvard University Press. Reprinted in *Philosophical Papers*, Volume I, pp. 122-130.

—. 1983a. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61:343–377, doi:10.1080/00048408312341131. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 8-55.

—. 1983b. *Philosophical Papers*, volume I. Oxford: Oxford University Press.

—. 1984. "Putnam's Paradox." *Australasian Journal of Philosophy* 62:221–236, doi:10.1080/00048408412340013. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 56-77.

—. 1986. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

—. 1991. *Parts of Classes*. Oxford: Blackwell.

—. 1992. "Meaning without Use: Reply to Hawthorne." *Australasian Journal of Philosophy* 70:106–110, doi:10.1080/00048408112340093. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 145-151.

—. 1994a. "Humean Supervenience Debugged." *Mind* 103:473–490. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 224-247.

—. 1994b. "Reduction of Mind." In Samuel Guttenplan (ed.), *A Companion to the Philosophy of Mind*, 412–431. Oxford: Blackwell, doi:10.1017/CBO9780511625343.019. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 291-324.

—. 1997. "Naming the Colours." *Australasian Journal of Philosophy* 75:325–42, doi:10.1080/00048409712347931. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 332-358.

—. 1999. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

—. 2000. *Papers in Ethics and Social Philosophy*. Cambridge: Cambridge University Press.

—. 2001. "Truthmaking and Difference-Making." *Noûs* 35:602–615.

Pryor, James. 2000. "The Sceptic and the Dogmatist." *Noûs* 34:517–549, doi:10.1111/0029-4624.00277.

Schwarz, Wolfgang. 2006. “Lewisian Meaning without Naturalness.” Draft, January 4, 2006. Downloaded from http://www.umsu.de/words/magnetism.pdf.

—. 2009. *David Lewis: Metaphysik und Analyse*. Paderborn: Mentis-Verlag.

Sider, Theodore. 2001a. “Criteria of Personal Identity and the Limits of Conceptual Analysis.” *Philosophical Perspectives* 15:189–209, doi:10.1111/0029-4624.35.s15.10.

—. 2001b. *Four-Dimensionalism*. Oxford: Oxford University Press.

Stalnaker, Robert. 2004. “Lewis on Intentionality.” *Australasian Journal of Philosophy* 82:199 – 212, doi:10.1080/713659796.

Weatherson, Brian. 2003. “What Good Are Counterexamples?” *Philosophical Studies* 115:1–31, doi:10.1023/A:1024961917413.

—. 2010. “Vagueness as Indeterminacy.” In Richard Dietz and Sebastiano Moruzzi (eds.), *Cuts and Clouds: Vaguenesss, its Nature and its Logic*, 77–90. Oxford: Oxford University Press.

Williams, J. Robert G. 2007. “Eligibility and Inscrutability.” *Philosophical Review* 116:361–399, doi:10.1215/00318108-2007-002.