Lewis on Reference and Eligibility

J.R.G.Williams

(October 24, 2011)

Contents

Section 1 gives an opinionated exegesis of Lewis' work on the foundations of reference—his *interpretationism*. I look at the way that, in the 80's, the metaphysical distinction between *natural* and *non-natural* properties came to play a central role in his thinking about language. Lewis's own deployment of this notion has implausible commitments, so in section 2 I consider variations and alternatives. In conclusion I outline a buck-passing strategy the Lewisian could adopt, either in combination or independently of the other maneuvers discussed. Section 3 briefly considers a buck-passing strategy involving fine-grained linguistic conventions.

1 Lewis's interpretationism

1.1 Interpretationism

David Lewis was no fan of primitive intentionality. He wanted to explain how the intentional—including mental and linguistic representation —could arise in a fundamentally physical world. He would agree, I think, with Hartry Field:

> there are no "ultimately semantic" facts or properties, i.e., no semantic facts or properties over and above the facts and properties of physics, chemistry, biology, neurophysiology, and those parts of psychology, sociology, and anthropology that can be expressed independently of semantic concepts. (Field, 1975, p.386)

Lewis's *interpretationism* is one implementation of a more general strategy for reducing the representational properties of language: the *select-and-project* method. Selection: an entire semantic theory is singled out based on facts about linguistic practice in a given population. Projection: *what it is* for N to refer to o (for that population) is for the selected semantic theory to entail that N refers to o.

Interpretationists like Lewis give a two-step account of selection. The first component is to specify (in terms of the community's linguistic practices) a 'target' pairing of sentences with semantic values. Lewis's own favoured version of the first step pairs sentences with (coarse-grained) truth-conditions; the pairing is established by appeal to conventional regularities in linguistic usage. Thus, if there is a convention among the relevant community of only uttering 'la neige est blanche' when they believe that snow is white, the pairing between that sentence and the set of situations where snow is white is established.[1] A different interpretationist account often associated with Lewis, *global descriptivism*, requires pairing sentences with truth *values* rather than truth *conditions*; perhaps every sentence that is treated by a community as 'platitudinous' is paired with the true; and every sentence that is rejected off-hand by the community is paired with the false.[2]

The second component moves from data to selected theory. A natural constraint is: to be selected, a semantic theory must assign semantic values to sentences that *Fit* the pairings on the target list. To keep things simple, let's understand 'Fit' in the most naive fashion: that the

[1] Notice that Lewis appeals to an intentional relation—belief—in formulating the target pairing. This is in line with his 'headfirst' methodology to the reduction of intention, whereby linguistic intentionality (reference, truth, etc) is reduced inter alia to mental intentionality (belief, desire, etc); a separate story of how the content of attitudes is fixed is promised. See Lewis (1994) for the headfirst/wordfirst contrast, and Lewis (1969, 1975) for an account of the relevant linguistic conventions.

[2] See Lewis (1983, 1984). A Putnamian alternative is to pair every sentence that is part of the idealized final total scientific theory with the true, and their negations with the false. Lewis is explicit that he is treating this global descriptivist theory as a simplified stalking horse, to illustrate and solve problems with his own more sophisticated account.

semantic theory predicts a pairing of sentences and contents that exactly matches those that appear on the target list.[3] One minimal account says that Fit is necessary and sufficient for selection; more elaborate accounts impose further selectional constraints.

In the early 80's, Lewis began advocating a constraint on the selection of semantic theory over and above Fit. He held that some theories were more 'eligible' than others to be selected, because they assigned to lexical items more 'natural' referents. This requires some explanation.

At the time he amended his interpretationism, Lewis had been convinced that for many tasks, throughout philosophy, an appeal to a distinction between *perfectly natural* and *merely abundant* properties was required. The distinction was more-or-less primitive. Perhaps a fan of Armstrongian Universals could explain it in terms of those properties that are necessarily coextensive with a Universal; perhaps we have a choice between positing naturalness as an unanalyzable division among properties and taking as basic a suitably rich contrastive resemblance relation among objects; but for present purposes this won't matter. Lewis had some specific proposals about what properties had this status. Fundamental physics, he thought, would be our best guide to the perfectly natural properties of the actual world.[4] He claimed, furthermore, that this all-or-nothing division between the perfectly natural properties and the rest, allowed us to make sense of a notion of relative naturalness—or more strongly, degrees of naturalness. The degree of naturalness of a property, says Lewis, is the minimal length of a definition of that property in perfectly natural terms.[5] Thus, for example, if having positive charge and mass of 1kg are each perfectly natural properties, having positive charge *and* unit mass might have a degree of naturalness of 1, where the degrees increase as naturalness decreases). For Lewis, semantic theories are graded as more or less *Eligible*, depending on the naturalness of the semantic values that they assign to lexical items—-the more natural (i.e. the shorter the definitional distance to the perfectly natural) the more eligible the theory.[6] Eligibility of semantic theory is a factor to be traded off against Fit.

Why bother with all this? A motivation for bringing in *something* other than fitting with the target data into one's story about how a semantic theory gets selected by linguistic usage is that, without it, the selection of semantic theory is simply too unconstrained. For example, suppose the Fit constraint on theory selection is that we generate the right pairing of sentences with coarse-grained truth conditions. Take a sensible seeming interpretation $\mathbb{I}$ generating the right pairing. It turns out we can construct a crazy seeming 'permuted variant' $\mathbb{I}^*$ which assigns the same truth conditions to sentences. Where the original theory said that 'est blanche' applied to white things, the permuted variant might say that something is in the extension of that predicate iff its image under the permutation ϕ was white. So we overgenerate (radically! drastically! absurdly!) selected theories. Our account of selection-by-usage hasn't filtered down to anything like the intuitively credible candidates to be the semantic theory of a language (still in the mix, for example, are interpretations on which my tokenings of 'that puffin' refers to a small furry creature orbitting Alpha Centauri).[7]

[3] See Lewis (1975) for discussion of all sorts of sophisticated elaborations of 'fit'—for example, taking pragmatic factors into account.

[4] The canonical statement of Lewis's views is in his (1983). Note that Lewis rejects any attempt at analyzing the natural in terms of fundamental science. In Lewis's eyes, this would require that physicalism be necessarily true; and while Lewis thought it was true, he held that non-physical perfectly natural properties were instantiated in other possible worlds.

[5] *other statements?*

[6] Though see also Lewis (1992), where the naturalness of the compositional axioms is appealed to. This will fit naturally into the perspective given below.

[7] For some of the original literature on this problem of the 'inscrutability of reference', see (Jeffrey, 1964; Quine, 1964; Davidson, 1979; Wallace, 1977; Putnam, 1981). See Williams (2008b) for a general version of the permutation argument in application to rich languages and for preserving sentential properties beyond truth values,

Lewis's Eligibility constraint speaks to this concern—for even though the original and permuted variant are equally fitting (assign the same truth values or propositions) prima facie the permuted variant will be less eligible than the original—it takes slightly longer to spell out in perfectly natural terms.

1.2 Reference magnets

We can understand the role the Eligibility constraint plays in terms of the concept of a *reference-magnetic* property. Rather than receiving a stipulative characterization within Lewis's framework, I'm going to treat this notion as functionally characterized. Suppose we have two semantic theories $\mathbb{I}, \mathbb{I}'$, differing only on the interpretation of one term t. $\mathbb{I}(t) = a$, $\mathbb{I}'(t) = b$. Suppose they would equally well explain linguistic usage. If t in fact refers to a rather than b, and furthermore in all similar situations where usage is equipoised, the symmetry is broken in the same way, then we say that a is reference-magnetic.

Independently of interpretationism, you can see in the permutation-style arguments given above a case for believing in reference-magnetism. The paradigmatic speech acts that constitute linguistic usage are performed with whole sentences: asserting, questioning, commanding and so forth. Their intended effects are characterized in terms of the truth-conditions of the sentences they involve (saying that the world meets those truth conditions, asking whether it does, commanding that it be made to do so). So—one may argue—interpretations that assign the same truth-conditions to whole sentences are 'equipoised with respect to usage' in the precise sense that they give the same input into speech-act theory.

But despite this equipoise, sensible interpretations get the facts about reference right and their permuted alternatives do not. So the symmetry over usage must be broken somehow. You needn't believe in reference-magnetism as characterized earlier to break the symmetry (you might instead appeal to some feature extrinsic to the candidate referents a and b—for example, a collective intention to use a term to pick out a particular object). But seeing reference-magnetism at work here is a natural thought, quite independently of Lewis's particular foundational theory of language.

There are more local ways to make the case for the existence of reference magnetism. Theoretical terms are a nice source of test cases. Often, we want to view old-time theorists as talking insightfully if sometimes false things about interesting subject-matters, in preference to taking them to be talking about less interesting subject-matters.Field (1973) makes a good case that the pre-20th century account of mass was equi-poised between rest mass, and the sum of mass and kinetic energy (what is sometimes called 'relativistic mass'). My informants tell me that unqualified 'mass' these days gets used for the former notion—it's regarded as more physically interesting, being independent of one's frame of reference, for example. Insofar as we want to say, in our present voice, that Newton talked and theorized about mass, we interpret him as talking about rest mass all along. While Field argued from equi-poise in usage to referential indeterminacy in old-timey deployments of 'mass', we will regard this as a case in which somehow the symmetry is broken and Newton's term ends up determinately picking out *mass*.

(This also illustrates a way in which reference-magnetism is independent of any particular foundational theory of language. For suppose one was a causal theorist of reference, and held that determinants of rest mass entered into causal relations, but determinants of the sum-of-rest-mass-and-kinetic-energy did not. Then it's very plausible that Newtonian usage of 'mass' picks out rest mass (since it's apt to be on the other end of reference-constituting causal relations). And the causal story will predict this instance of reference-magnetism).

and a discussion of why inscrutability is to be avoided for theoretical as well as intuitive reasons.

Lewis ties reference-magnetism to differences in Eligibility. If semantic theories are selected by Fit-plus-Eligibility, then when the former dimension is equipoised, the latter will break the symmetry. Given his account of what Eligibility consists in, entities will be reference-magnetic the more natural they are. This is a substantive claim, tying together a independently characterized concept with the minutiae of Lewis's particular views.

1.3 Reference magnetism and naturalness

The presentation above is deliberately sloppy in a couple of respects. First, I gave Lewis's characterization of degrees of naturalness in terms of 'definitional distance' of properties from the perfectly natural ones. But definitions are given in a certain language, and we don't pick out a language simply by listing a bunch of properties. Ted Sider in *Writing the Book of the World* suggests we respond to such concerns by extending the natural/non-natural distinction to entities of all categories, rather than just properties—in which case we can envisage a language 'Ontologese' whose every bit of vocabulary stands for something perfectly natural. Whatever the merits of this, it goes beyond the resources available to Lewis himself. I interpret Lewis as follows: we have a 'canonical' language, which might include conjunction, negation, unrestricted first-order universal quantification, identity, perhaps plural quantification and mereological overlap, and of course the usual variable and punctuational symbols. Let's call these collectively the 'auxiliary apparatus' of the canonical language. Aside from the auxiliary apparatus, the canonical language only contains predicates for perfectly natural properties. We can then say that P is definable in this language if there's an open sentence $\phi(v)$ of the language, such that necessarily, for all x, x has P iff x satisfies $\phi(v)$. However, semantic theories don't only assign semantic values to predicates; they also assign them to names, modifiers, operators, and so forth; and we may (indeed, I think we do) want to talk of the relative naturalness of the semantic values associated with these other kinds of lexical items. So we need an extended notion of definition to cover these cases. I'll assume we have one.

If we have the 'canonical language' laid down, and some favoured way of measuring the length of the definiens ϕ, then the notion of degree of naturalness can be taken to be the minimal length of such a definition—at least for those terms that have definitions at all. This is one locus for variation in a Lewis-style treat of eligibility, since it's not immediately obvious how lengths are to be measured, nor even what formal structure the 'degrees' will take—will they induce a total or partial ordering? Ordinal or cardinal? Lexiographic or Archimedian? Even if we don't want to commit to a full theory of relative naturalness at this stage, fixing on formal features of the ordering is important. I will assume that the lengths are measured by integer values, so the ordering is total, cardinal and archimedian (and with a natural zero). If you want a toy implementation, suggested by Lewis's writings, imagine that the length of ϕ is determined by counting the number of connectives that are present in ϕ. Notice that relations corresponding to the auxiliary apparatus—perhaps including identity and overlap—will be maximally natural, by this measure, even if they didn't appear on the list of perfectly natural properties and relations. Ultimately, these things will turn out to be 'reference magnets' by the lights of Lewis' theory, just as much as the natural properties are. (If you feel queasy about this, and are prepared to engage in the additional meaty metaphysics, you might consider the Siderian alternative).

But interpretationism does not appeal directly with relative naturalness. Instead *Eligibility*—which I'm using as a term for some sort of ranking of semantic theories—is the primary concern. This is a second locus for variation within the Lewisian account—for even assuming that it's only the naturalness of properties that matters, we're being asked to move from a ranking of individual properties, to a ranking of semantic theories that assign many properties, of various degrees of naturalness. I propose we think of the degree of eligibility of a theory as the sum of the degrees of naturalness of the semantic values it assigns in the lexicon; but notice that this only makes sense because we assumed that the initial comparative naturalness ranking assigned degrees that it makes sense to 'add together'. If we'd instead thought of the degrees as taking a partially-ordered structure, this would not be available to us. We'll come back to this point below.[8]

1.4 Best grammar and Humean simplicity

We have the Lewisian version of select-and-project metasemantics on the table. It uses conventional regularities in linguistic usage to identify a target pairing of sentences with propositions, and then selects the semantic theory that optimally trades off Fit with this data against Eligibility. The natural properties, because they make a semantic theory more eligible, are reference magnets. But presented in this way, the appeal to Eligibility comes in somewhat from left field. It (perhaps) solves some problems, and its reference-magnetic predictions may be attractive. But is there anything more to say about why it shows up in an account of language? I think the deeper story, from which Eligibility arises, turns on Lewis's handling of simplicity within his wider philosophy.

The appeal to natural properties only became a feature of Lewis's philosophy in the early 80's. But his account of language predates this. Here he is in 1975, responding to the puzzle we discussed above in connection to permutations—that fixing a pairing of sentences with propositions (which at this stage Lewis called a 'language') underdetermines subsentential reference (part of what Lewis called a 'grammar'). He has outlined a story about when a language counts as 'in use' in a given population, and turns to the question about grammar-selection:

> Unfortunately, I know no way of making objective sense of the assertion that a grammar Γ is used by a population *P*, whereas another grammar Γ' which generates the same language [translation: function from sentences to coarse-grained propositions] as Γ, is not. ...
>
> I do not propose to discard the notion of the meaning in *P* of a constituent or phrase, or the fine structure of the meaning of a sentence. To propose that would be absurd. But I hold that these notions depend on our ways of evaluating grammars, and are therefore no more objective than our notion of a *best* grammar for a given language.

Two key points to draw from this. (1) In 1975, Lewis's account of semantic theory ('grammar') selection, on the basis of sentence-level data ('language') is just that the selected theory is the best theory (where that is characterized by our standard evaluative methods). (2) He is worried that this story isn't 'objective'—in particular, that by tying grammar-selection to evaluative methods the former will be 'no more objective' than the latter.

[8] I've been writing as if the semantic theory proceeded by assigning semantic values to lexical items, and this seems natural if we think of something like a Lewis (1970) general semantics rather than, for example, a Davidsonian T-theoretic semantics (Larson & Ludlow, 1993). But even a general semantics can't do everything by the assignment of semantic values—Lewis's theory included the compositional axiom of function-application, and later advocated syncategoramic axioms governing lambda-operators. And it is arguable, I think, that even those sympathetic to possible worlds semantics should take a 'semantic theory' not simply to be an assignment function mapping expressions to semantic values, but rather an axiomatic theory that specifies such a function (see (Heim & Kratzer, 1998) for one version of what this would look like). There are natural adaptions of the above ideas to this setting—what takes the place of relative naturalness will be the length of the axiom governing a particular lexical item when formulated in the canonical language; and we again determine overall eligibility by adding this up.

Before moving on to tie this into eligibility, notice the immediate epistemic payoff of this early account of grammar-selection. The theorist of language arrives at judgements about what word mean by deploying certain evaluative methods. The truth about the subject-matter is fixed (on this account) by what the idealized application of those methods delivers. So there's no mystery about how the epistemic methods of the theorist are appropriate to their subject matter. Theories of selection that say anything else will generate an epistemic challenge—why are deployments of ordinary criteria good ways of finding out about semantic facts?[9]

One view of Lewis's development has him replacing or supplementing the earlier account of selection with a novel condition—naturalness now matters. This would mean that his later self would have to address the epistemic challenge. But I think this is the wrong interpretation. There is continuity here, which we can see by looking at how Lewis tackles parallel issues in his Humean theory of laws.

Lewis Humean account of laws of nature said that the laws were the generalizations entailed by a 'best' (optimally simple and informative) axiomatic theory. But the appeal to simplicity posed problems. It's a familiar point that the (syntactic) simplicity of theories in general depends on the language in which they're formulated. To give Lewis's example: if we have a primitive predicate F, that expresses 'being such that physical theory T holds', then the single axiom $\exists xFx$ will give us a maximally simple informational equivalent of T.[10] If by switching languages in this way we can achieve utter simplicity without sacrificing informativeness, then the simplicity component of the Humean theory loses traction, and the story collapses. In reaction, Lewis proposed that the relevant notion of simplicity operates on presentations of the theories concerned in a 'canonical' language, built out of the perfectly natural predicates—indeed, exactly the canonical language relevant to assessing relative naturalness. It is syntactic complexity in this privileged representation that matters for Humean system selection—and the introduction of artificial predicates F is neither here nor there. Let's call the notion of simplicity so characterized 'Humean'.

Combine the 1983 account of (Humean) simplicity with the 1975 view of semantic theory selection. Take an axiomatic presentation of semantic theory *in the canonical language*. The complexity of the axiom assigning a semantic value to an expression *in this presentation*, will, modulo some constant factor, be given by the complexity of the definition of that semantic value in canonical terms (what we called earlier the degree of naturalness of that semantic value). Adding the complexities of axioms together gives the overall Humean complexity of the semantic theory we start with. However, the same calculation (modulo some constant) gives the overall eligibility of the theory. So Humean simplicity/complexity and eligibility/(in)eligiblity measure the same thing.

Lewis's interpretationism remains constant in form throughout. It is projectivist in a full sense of the world, with the typical epistemic payoffs—it's just that the account of simplicity is elaborated. The reference magnetic behaviour of more natural properties, and the resolution of radical inscrutability threats this provides, thus drops out of the underlying story.

[9] As Paul Boghossian emphasized to me, the epistemological benefits of interpretationism depend on framing the epistemology in third-personal terms. The Lewisian story offer epistemic comfort to a theorist of language, who develops a theory of patterns of assent and dissent. But it's not so clear that ordinary ways of arriving at beliefs about what means what *in situ* as a language user are similarly vindicated. Presumably, an adequate story about language should be responsive to both, so even though Lewis's story gives a neat story about one half of the story, much remains to be done.

[10] At least, it does if we assume that informational equivalence is defined in coarse-grained terms. I don't want to pursue this further here.

2 Credible reference magnetism

The above discussion talked as if we get sensible results out of a Lewisian metasemantics, on which eligibility is analyzed ultimately in terms of definitional distance from the perfectly natural—which in Lewis's case, meant microphysical properties. But many feel that we are entitled to no such assumption. I don't really have a clue what a 'definition' of the ordinary subject matter of thought and talk—shoes and string and sealing wax—would be, if the definiens is to be drawn from microphysics. Further, I think that there are specific reasons to be worried that the account gives the wrong results—see Williams (2007a).

I won't argue directly against the Lewisian proposal here—so I'll leave it open for others to make the case that ordinary notions are finitely definable in the required way, and that the arguments that it leads to trouble can be blocked (perhaps by being appropriately subtle about some of the loci of variation in the account—the relation between perfect naturalness and relative naturalness, and relative naturalness and eligibility). Instead, I will work with the assumption that the Lewisian proposal as originally envisaged fails, and examine what prospects remain for an eligiblity-based interpretationism.

I'll look at three proposals: two that revise the overall metaphysics, and one that keeps the metaphysical framework intact but drops the connection between eligibility and naturalness.

2.1 Response 1: Macronaturalism.

The troubles for Lewisian eligibility, it might be thought, do not originate from the theory of eligibility per se, nor in its relation to perfect naturalness. The worries stems from the fact that Lewis commits to the view that (in the actual world) the perfectly natural properties are to be found in microphysics, and not in the 'macroworld' in which we operate. But, one might argue, wherever we find law-like connections; wherever we find genuine objective similarity; and wherever we find causation, we should believe that we're working with perfectly natural properties. And, it may be argued, we encounter such phenomena in the macroworld of geology, biology and ecology as much as the microworld of theoretical physics (Schaffer, 2004). The perfectly natural properties will be sparse but not ultra-sparse.

Whether eligibility-based interpretationism remains reductive in this setting is open to question—if we're allowing in the equivalent of biological and ecological Universals, what about psychological ones—beliefs and desires? Indeed, what about the special science of semantics? One might think that a principled version of this macroworld picture should include perfectly natural relations corresponding to intentional verbs, or even reference itself.

Even if we can take it that the vocabulary of the special sciences is available to us, it's not clear how we'd go about defining terms for artifactual kinds, nor the variety of verbs we use in everyday life (terms for thick ethical or aesthetic concepts, for example). So even if the range of resources we have available as a definitional base isn't as recherche as on the Lewisian proposal, the definitional ambition is still grandly ambitious. Of course, the reductive achievement in prospect, and the definitional ambition, play off one another: the more sparse a macroworld we buy into, the more reductive the final proposal, but the bigger the definitional task we set ourselves.

The macroworld view of the distribution of perfectly natural properties raises some questions about what the wider role of natural properties is to be. One idea that is prominent among Lewisians is that natural properties should enjoy a certain kind of modal independence. Chalmers' dualism (Chalmers, 1996) is a natural illustration of the sort of thing we might expect: on this kind of view, if some physical thing is conscious, it's possible for there to be a physical duplicate of it that is not a duplicate simpliciter, because it lacks consciousness. One question for the macronaturalist is whether something similar goes for their properties. Is it possible to have physical duplicates that are not chemical or biological duplicates, for example? If not, aren't we committed to some objectionable 'necessary connections between distinct properties'? Again, if special science kinds (for example) are perfectly natural, then won't the vagueness and indeterminacy of (e.g.) biological kinds give us 'vagueness in the world'?[11]

The macronaturalistic view demands we revise our entire conception of the fundamental structure of the world to support a metaphysics of one very special part of the world—linguistic representation. Is a more localized and modular response to our problems available?

2.2 Response 2: Comparative naturalism

A second response to worries about definability from microphysics is that Lewis went wrong (or that I went wrong in interpreting Lewis) by attempting to reduce relative naturalness to perfect naturalness via 'lengths of definitions' from some canonical language. Maybe we should stick with relative naturalness itself as primitive. After all, the motivating cases for this distinction are examples that involve comparative judgements very distant from what Lewis regards as perfectly natural: that green is more natural than grue; that *being the image under a permutation of something human* is less natural than being human; that artefactual kinds are less natural than biological kinds—and so forth. More general, the idea of a property P being 'grounded' or 'holding in virtue' of Q is fairly widespread in contemporary metaphysics, for better or worse—and one might think that this notion gives rise directly to comparative naturalness at all levels of reality. If comparative naturalness is rock-bottom, then an addition bonus (some may think) is that we can be agnostic over whether there is a layer of 'maximally natural' properties in the first place opens up—perhaps there are simply more and more natural properties ad infinitum (compare Schaffer, 2003; Langton & Lewis, 1998).

Just as with macronaturalism, we need to consider how primitive comparative naturalism integrates into wider theory. For example, it might sound attractive to countenance the possibility of ever-more-natural properties; but if the theoretical deployments of naturalness appeal to the all-or-nothing concept, then it's not clear that comparative naturalness will be an adequate replacement. As an illustration: Lewis's original theory of duplication and intrinsicality made appeal to the sharing of perfectly natural properties (cf. Lewis, 1983, 1986). The theory of Langton & Lewis (1998), which proved far more problematic, is one exactly designed to liberate the analysis of intrinsicality from appeal to a layer of maximally/perfectly natural properties. That it runs into worries that do not face the original account illustrates the damage such shifts in resources can inflict.

There are more local concerns about the adequacy of an appeal to primitive comparative naturalness in connection to metasemantics. Recall that earlier we emphasized the distinction between relative naturalness (of properties) and relative eligibility (of whole theories). Let's think of a toy case: an object language that has the syntax of first-order logic, with only three non-logical terms, the predicates 'rock', 'tree' and 'human'. To cut down on complexity, look at semantic theories that differ only over their assignments to the non-logical predicates. Compare three candidate assignments to 'rock', 'tree' and 'human', respectively: $A = (A_1, A_2, A_3)$, $B = (B_1, B_2, B_3)$ and $C = (C_1, C_2, C_3)$. We may suppose, writing $>$ for 'more natural than', we have: $A_1 > B_1 > C_1$, $B_2 > C_2 > A_2$ and $C_3 > A_3 > B_3$. This component-wise ranking gives us three different induced rankings of the overall assigned, which display a cyclic Condorcet-structure: A beats B twice, B beats C twice, and C beats A twice. In that kind of situation, how are we to

[11] For some exploration of life with a primitive macroworld, see Williams (2008c, 2007b). For vagueness in the world, see Williams (2008a); Barnes & Williams (2011).

rank the candidate assignments for relative eligibility?[12]

We earlier pointed out that one question for the theorist of comparative naturalness faced was over the formal characteristics of the ordering. Partial or total? Cardinal or ordinal? Lexiographic or Archimedian? A zero element or not? On some ways of answering this question, there will be natural ways of sorting out Condorcet-style troubles—for example, in the Lewisian account that (I took it) delivered integral degrees of naturalness, we could simply look at the sum total of the various degrees. But that kind of structure seemed reasonable because we were analyzing comparative naturalness in terms of a metrical notion—lengths of definitions. In the current context, there's no reason yet to think that summing 'degrees' of naturalness makes sense (particularly if they're partially ordered, or have no natural zero). And of course, the richer the structure we're forced to posit, the more costly it is to take it as explanatorily basic. The moral from these discussions is that we need to keep sharply in view that the resource that interpretationism needs is a ranking of whole theories, not individual properties. Without a story of how to get from the latter to the former, we have no theory at all.

2.3 Response 3: Parochial eligibility

The final option I will consider involves dropping the link between naturalness and eligibility, but retaining the rest of the structure of eligibility-based interpretationism. The story developed earlier can be seen as giving a programmatic specification of degrees of eligibility. We input some 'canonical language' L. We obtain an ordering of properties (and other entities) by minimum length of definitions in L—and by summing we thereby obtain what we might call 'degrees of eligibility$_L$' attaching to sets of entities/whole theories. Degrees of eligibility$_L$ are then traded off against fit to select the meaning-fixing theory of interpretationism. So each choice of L delivers an account of reference—reference$_L$. Question: for what L can we plausibly maintain that reference simpliciter is reference$_L$?

There are many candidates for L. Lewis (and others following him) propose some kind of construct drawn from fundamental Metaphysics ('Ontologese'), which we've seen causes problems. But L could be, for example, English. After all, the metalanguage of semantic textbooks tends to be a natural language (suitably supplemented by technical vocabulary)—not some artificially restricted language drawn from metaphysics or other special sciences. If we start with English, we have something with the recognizable pattern of eligibility-based interpretationism, but the base from which eligibility is determined is parochial rather than metaphysical.[13]

But if we choose the wrong L, the whole project could be threatened. Consider what one might say against the proposal of taking L=English (an extreme example, but a useful stalking horse). An overarching concern is that we are in effect offering a listiform, disjunctive characterization of what it is to be an eligibility-maker: x is an eligibility-maker iff x is either *being a shoe* or *being string* or *being sealing-wax* or... and so on throughout the English dictionary. But this seems an utterly ad hoc and arbitrary foundation to build a theory of reference on—why English rather than French or Mandarin? Why English-as-she-actually-is rather than English-as-she-might-well have been?

[12] Following up the analogy to voting paradoxes, it's worth asking whether a version of Arrow's theorem lurks in the vicinity. Pareto and non-dictatorship constraints are certainly plausible. However, it's not clear to me whether there's an interpretation of the Universality and Independence axioms that allows us to invoke the theorem in this setting. (Addendum: Hansen and Morreau, following up on the suggestion in this footnote, argue that a version of Arrow does indeed pose problems here).

[13] Compare **?**—Davidson's radical interpreters speak some particular natural language, and this kind of identification of the eligibility-basis introduces features reminisicient of Davidsonian radical intepretation into an otherwise Lewisian version.

There are a host of more specific worries. First, the ambitions of the appeal to eligibility might be unrealized. It was supposed to be reductive: but now reference is reduced in part to eligibility, and eligibility is defined, inter alia, via appeal to reference, truth and other semantic relations that English has words for. Moreover, part of the appeal of eligibility is that it was predictive: given the verdict that rest mass is more natural than rest-mass-plus kinetic energy, then we generate a prediction that an inchoate use of 'mass' picks out rest mass. But if English has a single lexical item 'mass', then all we get is the uninformative statement that the reference-magnet in the vicinity of 'mass' is mass (whatever that is!)

Second, the proposal looks extensionally inadequate. Suppose the French apply 'vert' roughly in the way English speakers apply 'green', but slightly shifted to the bluish end of the spectrum. Then it's very plausible that the extension of this colour term is slightly blue-shifted, compared to the English colour term 'green'. But if the properties for which we have English terms are reference-magnets, then our account would presumably predict that 'vert' picks out greenness—a false prediction. A related problem is that some words in English are so flexible that the whole account threatens to trivialize, leading us back to permutation problems or worse. If we can use the demonstrative 'that' to pick out some permuted variant of green, then a permuted interpretation of natural language could be phrased by simple demonstratives in the semantic axioms to assign a maximally 'eligible' permuted semantic theory.

Third, even if the account is extensionally adequate (getting the right actual meanings to actual words) it goes wrong in counterfactual scenarios. Consider the Ectoplasmians, living in an environment none of whose most important characteristics we lack words for. It seems possible for alien creatures in alien environments to refer to things around them. But there's no reason to believe an interpretationism parochially tied to the things *we* have words for as reference-magnets, will give decent results in such far-off scenarios.

I'm not going to investigate whether these obstacles can be overcome in any detail here—in particular because different choices of *L* face different versions of the challenges. But thinking through the worries about using English as *L* serves to: (a) highlight some of the benefits of Lewis's proposal that we may wish to preserve in a successor (reductiveness, preditiveness; absence of home-language imperialism; modal resource-sensitivity); and (b) indicate some of the hurdles that any successor should negotiate. For what it's worth, I think that one can come up with specific proposals for refining and improving (and disambiguating!) English to avoid several of the challenges just outlined. But there's a residue that any parochial choice of *L* must deal with. Of these, perhaps the most fundamental is whether the ad hoc nature of the choice of *L*, and the disjunctive character of the analysis of eligibility, rules the account out from the start.

Why should the ad hoc or disjunctive character of a proposed reduction detract from its credibility? After all, a disjunctive reduction is exactly what we should be looking for, if the property we're trying to analyze is itself disjunctive (consider grueness). Field (1972) famously compares the project of reducing reference to that of reducing chemical valence, and held that we should hold the latter to the same high standards (a listiform reduction of valence, he supposed, would be laughable). But he later qualified this position. What's bad about putative disjunctive reductions is that they failed to preserve explanatory role of property being reduced. Radically disjunctive properties aren't explanatory. If valence is explanatory, then we can reject a proposed identification with something disjunctive on the basis of Leibniz's law: valence has a property the proposed reductive basis lacks. For a similar argument to rebut a disjunctive characterization of reference, one would need to have confidence that reference has explanatory features that one can see that a mere disjunction lacks. Over time, Field lost confidence in this assumption: in recent times, he has been defending a sophisticated version of deflationism.

The same considerations play out in the interpretationist setting. If reference has a robust explanatory role, we should reject disjunctive specifications of it or the things to which it is reduced (eligibility). So the parochial proposals above will be no good. If reference has no explanatory role, then there's as yet no objection to it.[14] An interesting intermediate position arises if we think that *truth conditions* but not *reference* have an important explanatory role (the only contribution that facts about subsentential semantics make to explanation in general is via determining truth-conditions). Thoroughgoing deflationism about truth and reference would not be available by Fieldian criteria. But an interpretationism based on a robust account of conventions of truthfulness and trust, plus a deflationary account of eligibility, might remain a live option.

3 Buck-passing

Interpretationism has many virtues. Many attempts at a foundational account of language are incredibly limited in scope. They might have stories to offer about the reference of names of our friends and the predicates for things we bump into and trip over—but language contains far more. A total account of language needs to tell us about the meaning of modifiers, connectives, inflections, tenses, and all else. Interpretationism is a single-shot solution to such troubles. It also offers a tractable epistemology (at least for theorists of language)—so long as we stick within the broad constraints of the 1975 framework on which selected semantic theory is best semantic theory. The difficulty is to figure out how to understand this notion of 'best grammar'.

Up till now, I've been assuming (with Lewis) that the data to which the best grammar is responsible is a pairing of sentences with coarse-grained propositions. Even to get so far, Lewis has to endorse a 'headfirst' methodology whereby linguistic content is analyzed, not directly in terms of the non-intentional, but in terms that presuppose mental intentionality—the (coarse-grained) beliefs and desires in terms of which he characterizes what it is for a regularity to be conventional. So there's an element of buck-passing built into Lewis's account: propositional content is inherited from mental content via conventions; it is objectual content (and similar) that is the distinctive challenge at the linguistic level.

But there's nothing in the overall shape of Lewisian interpretationism that rules out a more thorough buck-passing. For example, unlike Lewis, we might help ourselves to fine-grained belief contents. We could have very fine-grained conventions—a convention regularity of assenting to 'that square is red' only when one has a belief whose content is the fine-grained proposition with first component a certain square, and second component redness. Such a rich set of sentence-proposition pairs would not allow much wriggle room for permutation style arguments.[15]

If one wishes to develop the account in this buck-passing direction (and still maintain the overall reductive ambitions), then there are a series of issues one must confront:

1. What kind of story is in prospect for reducing fine-grained mental content? Is the story interpretationist, or of some other kind? (Presumably it won't be convention-based interpretationist; but some kind of global descriptivism run on mentalese may yet be an option).

[14] Indeed, one interesting way of understanding deflationary eligibility is as a different implementation of the overall deflationary project. Deflationism itself has challenges of detail as well as of principle—for example, it's modal predictions at first pass seem whacky (if the specification of what reference is includes the fact that 'duck' refers to duck, it seems that 'duck' would still refer to duck even if our usage switched. Deflationists have to wheel in extra resources to explain this away (thus Field's Quinean appeal to translation). But interpretationism, deflating eligibility rather than reference, has less trouble.

[15] The reasons are much the same as in (**?**)

2. How rich is fine-grained mental content? Is it limited to e.g. combinations of macroscopic or phenomenal objects and properties? Does it allow boolean combinations thereof, quantification etc? Does it allow properties remote from direct experience to feature as constituents (electronhood)? Does it contain constituents corresponding to the features of natural languages (particularly those that vary): tense, aspect, case, modifiers, intensifiers, etc?

3. Is all fine-grained mental content reducible independently of linguistic content (and hence available for a non-circular reduction of the latter), or just basic part of it?

One thing that turns on these questions is whether there's any prospect of legitimately appealing to fine-grained conventions associating each sentence with what is in fact the structured proposition they express. That would appear to commit to an extreme view on which mental content can replicate all the structural complexities and idiosyncracies of natural language.

On the other extreme, it might be that fine-grained content is available, but in a very restricted form. We won't appeal to something like: assent to 'there's an electron' only when one believes that there's an electron present—since the fine-grained content is too sophisticated. Perhaps there are no conventional regularities associated legitimate fine-grained belief with assent-conditions. Or perhaps the sentence-proposition pairs that are delivered associate the sentence with its (macro/phenomenally articulated) verification conditions. But a semantic theory that aimed to connect sentences with those conditions wouldn't look much like what we find in contemporary textbooks!

When investigating such buckpassing variants on Lewisian interpretationism, bear in mind that the interpretationist story about 'Fit' *must* be fairly sophisticated, to cope with everday phenomenona such as metaphor, implicature, loose speech, exaggeration, and so forth. So even if a conventional regularity associates a sentence with something like a fine-grained description of default macroscopic/phenomenal verification conditions, a semantic theory that associates 'electron' with electrons may still count as 'fitting' that pairing. It would make the conventional regularity explicable given the background assumption that one asserts a sentence only when one has appropriate evidence that it's true.[16]

The buck-passing strategy is no panacea. There are several ways of generalizing the Lewisian appeal to conventions to exploit fine-grained content, and it's unclear which is best. It doesn't remove the need for hard thinking about constraints other than Fit in selection of best grammar (for example, the kind of weird compositional rules discussed in (Lewis, 1992)—and tackled by appeal to naturalness/eligibility—are still a threat). And it increases the demands on one's account of mental intentionality.

However, this might be a productive relocation of the problem. Nothing is a name by nature—the relationship between 'Aristotle' and Aristotle is paradigmatically extrinsic. Perception (for example) is very different. At first pass, it seems no coincidence that squarish percepts represent squares (hence the tradition, represented by Locke, on which the internal relation of resemblance is assigned a primary role in the analysis of intentionality). So in passing the buck, we don't just shift the problem of determining fine-grained content around—we relocate it to a position where both constraints and resources for tackling it are quite different.

Buck-passing interacts interestingly with some of the strategies from the previous section. I've already noted that something more than Fit will be needed in one's account of grammar-selection. But equally, a beefed-up Fit constraint might alter our evaluation of previous options. For example, parochial eligibility may seem more attractive, if constraints of Fit already remove a lot of the scope for arbitrariness.

[16] One is walking a fine line here, of course. A permuted semantic theory may make explicable fine-grained conventional regularities, given minimal competence in unwinding the permutations.

4 Conclusion

‘Eligibility’, and the idea of eligibility constraints on reference, has recently become rather vogueish. But it’s radically unclear, on reflection, what exactly is being referred to. One might stick with the letter of Lewis’s account—but then noone should have very high initial confidence that the theory is even extensionally adequate. One could go for one of the variants above—the superheavy metaphysics of macronaturalism, the seriously underdeveloped theory of comparative eligibility, or the defanged account of eligibility divorced from naturalness, that characterizes the parochial strategy. There are plenty more variations and alternatives! But appeal to eligibility without some indication about the intended framework is not a wise option for the conscientious reference-reducer.

References

Barnes, E. J., & Williams, J. Robert G. 2011. 'A theory of metaphysical indeterminacy'. *Oxford Studies in Metaphysics*, **6**.

Chalmers, David. 1996. *The Conscious Mind.* Oxford: Oxford University Press.

Davidson, Donald. 1979. 'The Inscrutability of Reference'. *The Southwestern Journal of Philosophy*, 7–19. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.227–242.

Field, Hartry H. 1972. 'Tarski's Theory of Truth'. *Journal of Philosophy*, **69**, 347–375. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 3-29.

Field, Hartry H. 1973. 'Theory change and the indeterminacy of reference'. *Journal of Philosophy*, **70**, 462–81. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 177-198.

Field, Hartry H. 1975. 'Conventionalism and Instrumentalism in Semantics'. *Noûs*, **9**, 375–405.

Field, Hartry H. 1994. 'Deflationist views of meaning and content'. *Mind*, **103**, 249–85. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 332-360.

Heim, Irene, & Kratzer, Angelika. 1998. *Semantics in generative grammar*. Oxford: Blackwell.

Jeffrey, Richard. 1964. 'Review of *Logic, Methodology and the Philosophy of Science*, ed. E. Nagel, P. Suppes and A. Tarski'. *Journal of Philosophy*, **61**, 79–88.

Joyce, James M. 1998. 'A non-pragmatic vindication of probabilism'. *Philosophy of Science*, **65**, 575–603.

Joyce, James M. 2009. 'Accuracy and coherence: prospects for an alethic epistemology of partial belief'. *Pages 263–297 of:* Huber, Franz, & Schmidt-Petri, Christoph (eds), *Degrees of belief.* Springer.

Langton, Rae, & Lewis, David. 1998. 'Defining 'intrinsic''. *Philosophy and Phenomenological Research*, **58**(2), 333–345.

Larson, Richard K., & Ludlow, Peter. 1993. 'Interpreted Logical forms'. *Synthese*, **95**, 305–356.

Lewis, David K. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, David K. 1970. 'General Semantics'. *Synthese*, **22**, 18–67. Reprinted with postscript in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 189–229.

Lewis, David K. 1975. 'Language and languages'. *Pages 3–35 of: Minnesota Studies in the Philosophy of Science*, vol. VII. University of Minnesota Press. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 163-88.

Lewis, David K. 1983. 'New Work for a Theory of Universals'. *Australasian Journal of Philosophy*, **61**, 343–377. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 8–55.

Lewis, David K. 1984. ‘Putnam’s paradox’. *Australasian Journal of Philosophy*, **62**(3), 221–36. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 56–77.

Lewis, David K. 1986. *On the Plurality of Worlds.* Oxford: Blackwell.

Lewis, David K. 1992. ‘Meaning without use: Reply to Hawthorne’. *Australasian Journal of Philosophy*, **70**, 106–110. Reprinted in Lewis, *Papers on Ethics and Social Philosophy* (Cambridge University Press, 1999) 145–151.

Lewis, David K. 1994. ‘Reduction of Mind’. *Pages 412–31 of:* Guttenplan, Samuel (ed), *A Companion to the Philosophy of Mind.* Oxford: Blackwell. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 291–324.

Putnam, Hilary. 1981. *Reason, Truth and History.* Cambridge: Cambridge University Press.

Quine, W. V. 1964. ‘Ontological Reduction and the world of numbers’. *Journal of Philosophy*, **61**. Reprinted with substantial changes in Quine, *The Ways of Paradox and Other Essays: Revised and enlarged edition* (Harvard University Press, Cambridge, MA and London, 1976) pp.212—220.

Schaffer, Jonathon. 2003. ‘Is there a fundamental level?’. *Noûs*, **37**, 498–517.

Schaffer, Jonathon. 2004. ‘Two conceptions of sparse properties’. *Pacific Philosophical Quarterly*, **85**, 92–102.

Wallace, J. 1977. ‘Only in the context of a sentence do words have any meaning’. *In:* French, P.A., & T.E. Uehling, Jr. (eds), *Midwest Studies in Philosophy 2: Studies in the Philosophy of Language.* Morris: University of Minnesota Press.

Williams, J. Robert G. 2007a. ‘Eligibility and inscrutability’. *Philosophical Review*, **116**(3), 361–399.

Williams, J. Robert G. 2007b. ‘The possibility of onion worlds’. *Australasian Journal of Philosophy*, **85**(2), 193–203.

Williams, J. Robert G. 2008a. ‘Multiple actualities and ontically vague identity’. *Philosophical Quarterly*, 134–154.

Williams, J. Robert G. 2008b. ‘The price of inscrutability’. *Nous*, **42**(4), 600–641.

Williams, J. Robert G. 2008c. ‘Working parts’. *In:* Le Poidevin, Robin (ed), *Being: Contemporary developments in metaphysics.* Royal Institute of Philosophy Supplement, vol. 83. Cambridge: Cambridge University Press.