Against Magnetism

Wolfgang Schwarz*

To appear in the *Australian Journal of Philosophy*

Abstract. Magnetism in meta-semantics is the view that the meaning of our words is determined in part by their use and in part by the objective naturalness of candidate meanings. This hypothesis is commonly attributed to David Lewis, and has been put to philosophical work by Brian Weatherson, Ted Sider and others. I argue that there is no evidence that Lewis ever endorsed the view, and that his actual account of language reveals good reasons against it.

1 The magnetic conception of meaning

How come 'tiger' means tiger and 'it is raining' that it is raining? Some say it's because of how these words are used in communication; some say it's because of descriptions or theories associated with the words; others say causal relations link the words to their meanings; and some argue that the nature of the meanings themselves plays a role. On this last account, certain things are by their very nature more eligible to be meanings than others – they are "reference magnets". The true meanings of our words are the ones that strike the best compromise between eligibility and whatever else enters into the determination of meaning. Call this *the magnetic conception of meaning*, or simply *magnetism*.

According to magnetism, what makes some candidate meanings more eligible than others is their objective naturalness, the fact that they carve nature at its joints. Tigers, for example, form a natural kind, and that is why the property *tiger* is a more eligible meaning than *tiger or elephant*, or *object that looks somewhat like a tiger*.

Magnetism has far-reaching philosophical consequences. Weatherson [2003] employs it to argue that knowledge may after all be justified true belief: if justified true belief is more natural than other candidate meanings, it could be the real meaning of 'knowledge' even it does not perfectly fit our intuitions and everyday usage. Along similar lines, Sider [2009] argues that even though nihilists and universalists about mereology systematically differ in their use of words like 'exists' or 'instantiated', we should not conclude that these words have different meanings in the two communities: perhaps the meanings that

* This paper is an updated and extended version of a manuscript originally circulated in 2006 under the title "Lewisian meaning without naturalness". Thanks to Karl Schaefer, Brian Weatherson, J. Robert G. Williams and the metasemantics reading group at the ANU for comments and discussion.

fit the nihilists' use are so magnetic that they are also the meanings for the universalists' words, despite the imperfect fit with use in that community.

The magnetic conception is often attributed to David Lewis. Indeed, Lewis writes in 'New Work for a Theory of Universals':

> 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 1983b: 47]

And in 'Putnam's Paradox':

> *Ceteris paribus*, an eligible interpretation is one that maximises the eligibility of referents overall. [...] [O]verall eligibility of reference is a matter of degree, making total theory come true is a matter of degree, the two desiderata trade off. The correct, 'intended' interpretations are the ones that strike the best balance. [Lewis 1984: 65f.]

However, if we look more closely at the context in which these quotes appear, and at Lewis's other works on language, I think it is clear that they do not reflect his actual views. On the contrary, we will see that Lewis's own position reveals good reasons *against* the magnetic conception of meaning.

In the next four sections, I will review the main tenets of Lewis's theory of mental and linguistic content, and discuss how considerations of objective naturalness fit into this picture. Then I will explain why Lewis appears to defend a rather different account in 'Putnam's Paradox' and pp.45–49 of 'New Work for a Theory of Universals'. In section 6, I take a step back and argue that magnetism not only conflicts with Lewis's particular views on meaning, but with the simple idea that semantics should have a place in a more comprehensive theory of rational communication and human interaction. In section 7, I argue that certain observations that appear to support magnetism do not really support it.

2 Lewis on language

Lewis's views on meta-semantics were set out in his first book *Convention* [1969], and underwent only minor changes thereafter (see especially [1975], [1979b], [1980a], [1986b: 40–50] and [1992][1]). In a nutshell, Lewis held that sentences have their meaning in virtue of their use in a linguistic community. For the kind of sentences that concern us here, the relevant patterns of use largely concern the conditions under which a sentence may be uttered by the conventions of the community. For example, in the German-speaking

1 'New Work for a Theory of Universals' [1983b] and 'Putnam's Paradox' [1984] are not on this list, for reasons given in section 5.

community, there is a convention to utter 'es regnet' only when one believes that it is raining. According to Lewis, a convention is first and foremost a regularity. People generally do utter 'es regnet' only when they believe that it is raining. Moreover, the existence of the regularity is common knowledge in the community. That is why people who hear an utterance of 'es regnet' can infer that the speaker probably believes that it is raining, and also, if the speaker is deemed reliable in these matters, that it is probably raining. Although the regularity is in a sense arbitrary – one could have used other sounds or marks instead of 'es regnet' – it is useful to the community and sustained by shared interests of its members.

The convention governing the use of 'es regnet' among speakers of German links the sentence with a condition on possible situations. Roughly speaking, this is the condition that must be satisfied (or believed to be satisfied) for an utterance of the sentence to be conventionally appropriate. For 'es regnet', the condition filters out the possible situations at which it is raining. Other sentences are paired with other conditions. Let us say that a sentence is *true* at a possible situation (relative to a given community) iff the situation satisfies the condition conventionally associated with the sentence. The task of a *grammar*, as Lewis uses the term, is to offer simple rules that generate this association between sentences and truth-conditions, based on semantic values assigned to a finite set of lexical items. 'To plug into its socket in an account of the use of a language, a [...] grammar has to specify which speakers at which times at which worlds are in a position to utter which sentences truly' [Lewis 1986b: 41].[2]

In this way, Lewis's meta-semantics connects the powerful tradition of model-theoretic semantics with the intriguing but systematically rather less fruitful idea of meaning as use. The bridge is effected by functions mapping sentences to truth-conditions. These abstract objects, which Lewis calls *languages*, are the topic of model-theoretic semantics, which can therefore look like a branch of mathematics. However, to say that a language, in this sense, is the language of a given community, is to say no more (and no less) than that there obtains in the community a certain complex system of social conventions. Lewis's theory also manages to integrate the Gricean tradition where meaning is analysed in terms of speaker intentions (see [Grice 1957]): Lewis [1969: 155] shows that under favourable circumstances, the presence of a linguistic convention entails that when a speaker utters a sentence associated with a condition p, then she typically tries to get the addressee to realise that she (the speaker) wants him (the addressee) to come to believe that p.

2 This is obviously only a rough sketch of Lewis's account. Among other things, I have ignored non-assertoric speech acts, implicatures, vagueness, and the "two-dimensionalist" distinction between truth-conditions in the present sense ('meaning$_2$' in [Lewis 1975: 173]) and truth-conditions in the sense of Kaplanian *content* ('meaning$_1$'). My sketch matches what Lewis himself took to be the core of his meta-semantics (e.g. in [Lewis 1975], [Lewis 1986b: 40–50] and [Lewis 1992]), and the complications would not make a significant difference to the topic of this paper.

The job of semantics, in this picture, is to assign the right truth-conditions to sentences – the truth-conditions that fit the patterns of use. Notice what is *not* part of that job description. First, a semantics need not model the cognitive mechanisms underlying speakers' linguistic competence (see [Lewis 1969: 183, 199f.], [Lewis 1975: 178], [Lewis 1980a: 24], [Lewis 1992: 151, fn.6]). Second, the semantic values employed in a grammar need not qualify as *referents* or *meanings* in any traditional or pre-theoretic sense of these terms. 'The object', Lewis explains, 'is not that we should find entities capable of deserving names from the established jargon of semantics' [Lewis 1986b: 41]. That is why he generally avoids speaking of 'reference', 'meaning' and even 'intension', and prefers the neutral 'semantic value' (see [Lewis 1974b], [Lewis 1980a: 25f.], [Lewis 1986b: 41f.]).

Consider reference. No doubt 'London' refers to London. Our concept of reference is tied to the disquotation schema "'x' refers to x". Any theory according to which common instances of the schema are false, or not determinately true, is unacceptable – it is simply not a theory of *reference*. But we cannot take for granted that our intuitive concept of reference plays a role in a systematic grammar for English. Following [Burge 1973], it has been suggested that names like 'London' are fundamentally verb phrases, whose semantic value is something like a function from worlds to extensions. Relatedly, Lewis [1970a: 216f.] and Montague [1973] discussed the idea that names should be treated as quantifiers, with sets of sets as semantic values. "Semantic values may be anything, so long as their job gets done" [Lewis 1980a: 26].

To what extent do our linguistic conventions – the facts about use – determine a unique language, a unique pairing of sentences with truth-conditions? It is sometimes suggested (e.g. in [Sider 2009: 400]) that facts about use merely select a set of sentences which an interpretation must render true. This would radically underdetermine the truth-conditions of every sentence: it would leave open whether 'es regnet' means that it's raining or that 2+2=4. For Lewis, use does much more than single out a set of sentences, i.e. an association between sentences and truth-values. Linguistic conventions effectively associate sentences with full-fledged truth-conditions. The hypothesis that 'es regnet' means that 2+2=4 is incompatible with various facts about use – for instance, with the fact that people uttering the sentence typically intend to convey that it is raining, not that 2+2=4.

Nevertheless, it is plausible that linguistic conventions do not select *unique* pairing of sentences with truth-conditions. For one thing, competent speakers are often uncertain and in disagreement about the correct usage of a word, especially when applied to unusual cases or in unusual circumstances (see [Lewis 1975: 188]). In addition, our usage may not settle the truth-conditions for very long and complicated sentences that nobody ever utters (see [Lewis 1992]).

The underdetermination gets more severe when we turn from "languages" (pairings of sentences with truth-conditions) to "grammars". Any mapping from sentences to truth-

conditions can be generated by infinitely many grammars, most of them incomprehensibly complex and unintuitive. That is why the task of semantics, for Lewis, is not just to come up with any old grammar that delivers the right truth-conditions, but with a *systematic*, *simple* and *intuitive* grammar (see [Lewis 1969: 198], [Lewis 1975: 176f.]). Lewis does not explain these virtues in any detail. One might, for example, require that semantic values assigned to individual words should preferably bear some relation to what we ordinarily call the referent or meaning of the words: a grammar that identifies the meaning of 'London' with London is *ceteris paribus* better than one that identifies it with the number 7.

At this point, we could stipulate that the true grammar among those generating the right truth-conditions is the one with the objectively most natural semantic values. This idea is sometimes attributed to Lewis, based a few passages in [Lewis 1984] taken out of context. Technically, the proposal might count as a form of magnetism, although the label is a bit misleading, as the force exerted by natural meanings would be independent of distance (i.e., of similarity). We would also not get the philosophically interesting consequences exploited by Weatherson and Sider, since facts about naturalness would never affect the truth-value of any sentence. I will therefore set aside this kind of magnetism. For what it's worth, I think it would fit badly into the Lewisian approach. If we take semantics to be part of a systematic study of human interactions, why should we want the semantic values that figure in grammars to be objectively natural? Given the choice between an elegant and intuitive grammar whose semantic values are objectively somewhat unnatural and a less elegant and intuitive grammar whose values are objectively more natural, would we really judge that the second is better? Why not use the first, since it is simpler and also generates the right truth-conditions – which is the ultimate goal?

As far as I can tell, objective naturalness plays essentially no role in Lewis's theory of language. Lewis had fully developed his views before he began to believe in objective naturalness in the early 1980s (see [Lewis 1999: 1]), and there is no sign that he later changed his mind. Naturalness is still not mentioned at all in the summary of his philosophy of language in [Lewis 1986b: 40–50]. On the other hand, Lewis did come to believe that objective naturalness matters to the content of mental states. Since the meaning of sentences depends on the attitudes in the linguistic community, considerations of naturalness thereby affect, albeit indirectly, the meanings of sentences. We thus need to turn to Lewis's theory of mind. (Central texts here are [Lewis 1974a], [Lewis 1979a], [Lewis 1980b], [Lewis 1986b: 27–40] and [Lewis 1994].)

3 Lewis on intentionality

Lewis's convention-based account of language is not very useful if we wonder how our beliefs and desires get their content. Here Lewis appeals to a more traditional form of functionalism: whether a state is a belief that it's raining, a desire for soup or an intention to only utter 'es regnet' when it is raining, is largely determined by the state's causal profile, its typical causes and effects. Which causal profile goes with which state is registered by "folk psychology", our common but tacit understanding of the interactions between beliefs, desires, perceptions and rational action.

According to Lewis, folk psychology looks 'a lot like Bayesian decision theory' [Lewis 1979a: 148f.]. Bayesian decision theory attributes to agents a pair of a probability function P and a utility function U, representing the agent's degrees of belief and desire, respectively. In any given situation, a Bayesian agent chooses whichever act has the highest expected utility by the lights of P and U. Turning this around, we may therefore attempt to read off an agent's probabilities and utilities from their actual and counterfactual choices.

So-called representation theorems in decision theory seem to show that if an agent's choice dispositions satisfy certain qualitative constraints, then there is *unique* system of beliefs and desires that matches her dispositions (see e.g. [Savage 1954]). Lewis did not trust these results and argued that, on the contrary, our choice dispositions leave some aspects of our attitudes radically underdetermined (see especially [Lewis 1983b: 50–52]). The disagreement turns on rather subtle issues about the calculation of expected utility and the definition of acts and outcomes. Without going into the details, I will assume that Lewis was right on this matter.[3]

To fix the aspects of content that are left open by choice dispositions, Lewis turns to other compartments of folk psychology, concerning not the output of mental states, but their input. "Folk psychology says that beliefs change constantly under the impact of perceptual evidence: we keep picking up new beliefs, mostly true, about our perceptual surroundings" [Lewis 1994: 320]. Remarks like this frequently occur in Lewis's writings (see [1980c: 274], [1983a: 380], [1983b: 50], [1986b: 106], [1994: 299f.], [1997: 334f.]), but they are never developed in any detail. The following picture, this time taken from Bayesian epistemology, seems to fit what he had in mind.

Perceptual experiences impose a constraint on the agent's beliefs. In simple (if somewhat idealised, see [Lewis 1986d]) cases, the constraint says that a certain proposition, capturing

3 Lewis's argument in [Lewis 1983b] uses Jeffrey's [1965] "evidential" notion of expected utility. It is known that in Jeffrey's decision theory, unlike in Savage's, hypothetical choices (or preferences) do not determine a unique assignment of beliefs and desires (see [Joyce 1999: 122–127]). Lewis himself endorsed a version of Savage's theory (see [Lewis 1981]). However, he presumably rejected Savage's permissive conception of acts as arbitrary functions from "states" to "outcomes", which is crucial for the proof of Savage's representation theorem.

the agent's "evidence", should receive probability 1. Here a proposition should not be understood as a sentence, which would be in need of interpretation, but as a way things might by – i.e. a property, or a set of centred worlds (see [Lewis 1979a]). Since our perceptual evidence is supposed to reflect facts about the actual perceived environment, we can assume that it rules out possibilities in which the environment is in certain respects different. For example, when you look at a black raven, your visual evidence might rule out possibilities in which you are confronting, under otherwise ordinary circumstances, a pink elephant.[4] To take into account this new evidence (call it E), your probability distribution now changes by conditionalisation, so that your new degree of belief in any proposition A equals your previous conditional degree of belief in A given E.

This ensures that you assign probability 1 to any proposition logically entailed by the evidence. The effect on most other propositions depends on your previous probabilities. If you are convinced all along that the ravens you get to see have untypical colours, then the observation of black ravens might increase your confidence that most ravens are white. Lewis therefore holds that folk psychology also constrains what kinds of evidence an agent may take to support what hypotheses. In our simple Bayesian model, this can be expressed as a constraint on the agent's *ultimate priors* – a hypothetical probability function representing the agent's beliefs before taking into account any evidence. For example, folk psychology might say that under normal circumstances, people take the observation of black ravens to support the hypothesis that unobserved ravens are black, and not that they are white. This means that agents assign relatively high prior probability to situations in which observed and unobserved ravens generally agree in colour.

This constraint on prior probabilities belongs to what Lewis, following [Grandy 1973], calls *principles of humanity* (see [Lewis 1974a: 112f.], [Lewis 1983b: 52–54], [Lewis 1986b: 38f., 107], [Lewis 1994: 320].) Other principles of humanity constrain an agent's basic desires. To use Lewis's favourite example, they might say that people do not have a basic desire for a saucer of mud.

In a similar context, Davidson suggests a kind of master principle that underlies all principles of humanity: we should interpret others as sharing *our own* basic beliefs and desires (see e.g. [Davidson 2004]). The individual principles of humanity follow from this master principle together with facts about the contents of our own beliefs and desires. Lewis could not follow Davidson here, since he rejects Davidson's interpretationist account of attitudes. For Lewis, the content of an agent's beliefs and desires is an objective matter, determined by their non-intentional properties; the actual or hypothetical presence of an interpreter is completely irrelevant. Hence Lewis cannot appeal to attitudes of the interpreter in his account of intentionality. Even if he could, this would not meet his

4 Here I ignore Lewis's puzzling suggestion in [Lewis 1996] that our evidence consists of propositions about our brain states.

reductionist aspirations.

Lewis does not offer a single, unified principle of humanity. However, in the early 1980s he came to believe that *some* of the principles derive from a general constraint to the effect that the content of attitudes must have a reasonably high degree of objective naturalness. This is where objective naturalness enters Lewis's theory of mind and language. Before I discuss Lewis's proposal, let us have a quick look at Lewis's (sometimes misunderstood) views on objective naturalness.

4 Lewis on naturalness

To understand Lewis's conception of natural properties it is perhaps best to start with the idea of a perfect duplicate. Imagine someone creates a perfect replica of the *Mona Lisa* that not only looks like the original to untrained eyes, but cannot be told apart from the original by even the best of forensic equipment. Despite their similarities, the Mona Lisa and its duplicate differ in infinitely many respects – for example, with regard to membership in the set { the Mona Lisa, the Eiffel tower }. If all properties are on a par, there is nothing distinctive about the pair of Mona Lisa and its duplicate as compared to, say, the pair of Mona Lisa and the Eiffel tower. But surely there is an important difference between these pairs, a difference that is not created by our contingent interests or habits. Classes of perfect duplicates are *objectively* distinguished from arbitrary other classes of objects.

What lies behind this objective distinction? What does it take for something to be a perfect duplicate of the Mona Lisa? Intuitively, original and duplicate must be composed of exactly the same (types of) particles, arranged in exactly the same way. Physicists characterise fundamental particles by a number of basic features such as mass, charge and spin. Taken together, these features divide the particles into classes of perfect duplicates. Like the Mona Lisa and its replica, any two electrons are perfect duplicates of one another, in virtue of having the exact same mass, charge, spin, etc.

Considerations like these lead Lewis to the following picture (see especially [Lewis 1983b], [Lewis 1986b: 59–69], [Lewis 1994: 291–297], [Lewis 1998b], [Lewis 2009]). There is a special class of *fundamental properties and relation* (perhaps including mass, charge, spin and spatiotemporal distance) that figure in the basic laws of physics and whose distribution determines whether two objects are perfect duplicates: an object A is a perfect duplicate of an object B iff there is a one-one map between their atomic parts that preserves fundamental properties and relations. This also applies to possible worlds. Two worlds are perfect duplicates iff they agree in the distribution of fundamental properties and relations. Thus all (qualitative) truths about any world supervene on the distribution of fundamental properties and relations at that world: once the distribution of fundamental properties and relations is fixed, the whole world is fixed. Moreover,

the fundamental properties and relations provide a *minimal* basis: if you specify the distribution of some, but not all, fundamental properties, you have not yet fully specified a world. Although the laws of physics state connections between fundamental properties, these connections are metaphysically contingent. Across logical space, fundamental properties and relations are freely recombinable.

The fundamental properties and relations Lewis also calls *perfectly natural*. This kind of fundamentality must not be confused with conceptual or ideological fundamentality. Mass, charge and spin are good candidates for fundamental properties, but our words or concepts for these magnitudes are hardly primitive. For Lewis, they would be paradigmatic examples of theoretical terms, defined by their role in physical theories (see [Lewis 1970b]). In the other direction, notions like identity, parthood and fundamentality may count as undefined primitives in Lewis's metaphysics, but they fit no part of the job description for fundamental properties. Identity, for instance, is not at all freely recombinable: it can never obtain between different objects, and it can never fail to obtain between an object and itself. At the beginning of 'New Work for a Theory of Universals', where Lewis gives a list of properties that are definitely *not* fundamental, he mentions, besides things like *grue*, also identity, parthood, set membership and fundamentality (or rather, similarity in a basic respect, which for Lewis is interdefinable with fundamentality). This was certainly not a mistake.[5] Sider's [2011] recent proposal to extend Lewis's conception of fundamentality to logical notions like negation, identity and existence is therefore distinctly un-Lewisian and presupposes a very different conception of metaphysics.

The label 'perfectly natural' suggests that there are also *im*perfectly natural properties. The idea is intuitive. Actual copies of the Mona Lisa are not perfect duplicates, but they still share an objective similarity with the original. Properties like solidity or radioactivity make for objective similarity, but they are hardly fundamental. Lewis held that a property's degree of naturalness is somehow determined by its connection to perfectly natural properties, but he never gave a satisfactory account of how. In [1986b: 61] and [1984: 66], he suggests that what matters is length of definition in terms of perfectly natural properties. (If definitions are understood as Boolean combinations, this would have the undesired consequence that extrinsic properties or functional properties all come out as maximally unnatural.) In [1983b: 49], he suggests that a property's degree of naturalness is related to the extent to which its instances are distinguished from their surroundings in terms of fundamental properties. Earlier in the same paper (pp.13f.), he suggests that degrees of naturalness might somehow be analysed in terms of similarity between universals (which would require overcoming the problems raised in [Lewis 1986a]). Perhaps he took these ideas to be compatible. It is also worth keeping

5 Admittedly, at other places Lewis is not as clear about the distinction between fundamentality and undefinability as he should have been. The importance of this distinction is emphasised in [Busse 2009].

in mind that different measures of naturalness might be useful for different purposes. Notably, by the criteria described in [Lewis 2001] the perfectly natural properties come out as relatively *unnatural*.

Now suppose we have somehow defined an objective measure of naturalness for non-fundamental properties in terms of fundamental properties that privileges, say, *green* over *grue*. We might then appeal to this measure (or one such measure) in a theory of mental content. This is what Lewis suggests in 'New Work for a Theory of Universals'. Again the suggestion is only presented as a brief sketch. Since Lewis does not assume that beliefs and desires are composed of word-like parts, we can't require that these parts should have natural referents. The bearer of content is an entire brain state, and its content is given by something like a probability and utility distribution. But remember that these distributions assign probabilities and utilities to properties. Roughly speaking, an agent assigns high probability to a property p iff she is confident that she herself instantiates p. Lewis seems to suggest that agents assign high probability (and utility?) only to reasonably natural properties.

This does not look very promising. First of all, the rules of probability dictate that if p has high probability and q has any probability at all (even zero), then the probability of p *or* q is at least as high as that of p, not matter how gerrymandered this proposition. Secondly, as Weatherson [Manuscript] points out, even if *green* is more natural than *grue*, it is not at all obvious that *being in a world where all emeralds are green* is objectively more natural than *being in a world where all emeralds are grue*. (Although the issue is hard to adjudicate, without a concrete measure of relative naturalness.) Third, for many of our attitudes, considerations of naturalness that trace back to distinctions in fundamental physics just seem out of place. A lot of people desire peace, wealth, or holidays on the Bahamas. I do not know how natural or unnatural these properties are by the lights of fundamental physics, nor how natural they are compared to other possible objects desire. But I do know that it does not matter. Folk psychology does not make the attribution of such desires conditional on microphysical discoveries about objective joints in nature.

But there are other, more plausible ways of spelling out Lewis's idea. Return to the simple Bayesian model from section 3. Here we used principles of humanity to constrain an agent's prior probabilities and thereby their attitudes about evidential support. For example, rational agents should assign high prior probability to the assumption that nature is uniform. But uniform in what respect: should one believe that emeralds are uniform with respect to *green* or with respect to *grue*? Here we can appeal to natural properties: one should assign high probability to worlds that are uniform with respect to patterns in the distribution of fundamental properties. At least in worlds like ours, attributes like *green* (unlike *grue*) supervene on intrinsic physical features of their instances: perfect duplicates never differ in colour. Hence if unobserved emeralds are

similar to observed ones in their fundamental physical properties, it is plausible that they will also be green and not blue. It does not matter, for this proposal, whether *green*, or *being in a world where all emeralds are green*, are themselves particularly natural.

Lewis saw further opportunities to use considerations of naturalness in his principles of humanity. On pp.53f. of 'New Work', he discusses the Kripkensteinian challenge to explain why, when we do arithmetic, we intend to add rather than quadd. His answer is that the activity of adding is more natural, and therefore more apt to be the content of an intention, than the activity of quadding.[6] His discussion of deviant grammars in [Lewis 1992] can be read as following the same line of thought: let Quenglish be a language that differs from English only in the truth-conditions assigned to very long and complicated sentences that nobody ever uses. In virtue of what do we speak English rather than Quenglish? Perhaps it is because speaking English is objectively more natural than speaking Quenglish. In either case, Lewis seems to assume that the greater simplicity of an abstract rule makes following the rule a more natural activity, and therefore a better candidate to be the object of attitudes. It is not clear to me that this kind of naturalness bears any connection to fundamental properties, or that such a connection would be needed for a response to Kripkenstein, but that is what Lewis seemed to think (at least in [Lewis 1983b]; objective naturalness is never mentioned in [Lewis 1992]).

Setting aside these matters of detail, let us look at the difference Lewis's new-found belief in objectively natural properties made to his theory of mind and language. The difference lies entirely in the formulation of the principles of humanity. Previously, the principles said that an agent's priors should take the world to be uniform in certain specified respects: with respect to the greenness of emeralds, perhaps, or with respect to patterns in the distribution of mass, charge and spin. Now these constraints are unified by a single constraint: an agent should assume uniformity in the distribution of fundamental properties and relations. Similarly, where we previously might have had a constraint that privileges intentions to add over intentions to quadd, this is now (supposedly) derived from a general constraint that privileges intentions to perform natural activities.

Not all principles of humanity follow from these constraints. Irrationally skeptical beliefs, and irrational responses to evidence, are not always based on counterinductive priors. For example, rationality also demands some kind of open-mindedness or information-theoretic entropy. Lewis's constraints on basic desires also have little to do with objective naturalness. A saucer of mud is hardly less natural in any objective sense than a saucer of soup, or world peace.

If we follow Lewis and assume that, *via* principles of humanity, there is a naturalness constraint on mental content, this has consequences for the truth-conditions of sentences.

6 Contrary to widespread reports, Lewis never suggested that the abstract mathematical function of addition is objectively more natural than quaddition. Sider [2004: 15] cites this as a precedent for extending naturalness to logical and mathematical notions.

'Green' means *green* in part because people tend to communicate beliefs about greenness when using 'green', and their beliefs are about greenness rather than grueness in part due to the greater naturalness of *green*, or perhaps due to the fact that fundamentally similar objects generally agree with respect to *green* and not *grue*. The result looks a bit like magnetism. This might explain why, at two places in his writings, Lewis appears to endorse the magnetic conception of meaning: in 'Putnam's Paradox' [1984] and on pp.45–49 of 'New Work for a Theory of Universals' [1983b].[7] But if we look more closely, it is clear that Lewis did not actually believe what he was writing there.

5 Global Descriptivism and Putnam's Paradox

To follow their linguistic conventions, the members of a community need to know, at least implicitly, under what conditions their sentences may be truthfully uttered. These conventional truth-conditions do not always reflect how speakers evaluate a sentence relative to counterfactual scenarios. Perhaps Queen Elizabeth II is essentially the daughter of George VI, in which case there is no counterfactual situation at which Queen Elizabeth is not the daughter of George VI. Nevertheless, there plausibly are conditions under which it would be correct by our linguistic conventions to deny that Queen Elizabeth is the daughter of George VI. For instance, we might discover that Queen Elizabeth is actually the daughter of Winston Churchill. It looks like the information conventionally associated with utterances of 'Queen Elizabeth is so-and-so' does not select the class of possible situations at which Queen Elizabeth herself has a certain property, but (roughly) the class of situations at which whoever constitutionally reigns over the UK throughout the late 20th century and stands in a suitable connection to our use of the name 'Queen Elizabeth' has that property.

This is how descriptivism fits into Lewis's theory of language. It is not an alternative to the convention-based semantics, but a part of it. In this form of descriptivism, what people associate with linguistic expressions are *conditions*, i.e. properties, not further expressions (see also [Jackson 1998b: 201–204]). That we happen to speak a fairly rich language in which we can express the truth-conditions of many sentences by means of different sentences is, linguistically, a mere coincidence. However, it is a coincidence that makes it possible to wonder how things described in one fragment of our language relate to things described in another fragment – how, say, moral or psychological truths relate to facts expressed in physical vocabulary. This is why in Lewis's work, descriptivism mainly comes up in metaphysics (like [Lewis 1966], [Lewis 1989], [Lewis 1994]), not in

7 Footnote 6 of [Lewis 1993] might also be cited, although it only concerns thought, not language. In footnote 2 of [Lewis 2004], Lewis mentions the hypothesis that even though there are many candidate referents for paradigmatically vague terms like 'the cloud', one of these candidates is the true referent in virtue of being 'a mighty reference magnet', but he does not appear to endorse this idea.

the philosophy of language. (Accordingly, 'How to define theoretical terms' [1970b] is in the 'Ontology' section of his *Philosophical Papers I*, not in the 'Language' section.)

In 'Putnam's Paradox' [1984], Lewis discusses a different account on which our words do get their meaning by the linguistic descriptions, or theories, in which they figure. The idea is that the correct interpretation of our words is the one that makes our total theory come true. This Lewis calls *Global Descriptivism*.

Global Descriptivism has seemed attractive to some, perhaps because it promises to cash out the idea that all words ultimately owe their meaning to some kind of definition. But on closer inspection, it is a really bad theory. Let me mention just three of its most obvious problems.

First, Global Descriptivism ignores many facts about sentences that are clearly relevant to their meaning: under what kinds of circumstances they are typically uttered, what purpose is served by their utterance, how they affect the hearer's beliefs, and so on.

Second, Global Descriptivism makes expressive redundancy a precondition on meaningful languages, which is absurd. Consider a simple language with only three sentences, used to convey the current state of the weather: 'it is raining', 'it is snowing' and 'it is sunny'. None of the terms in this language are definable by means of any others, but certainly not any interpretation that assigns something true to one of the sentences and something arbitrary to the others is correct.

Third, without further constraints, Global Descriptivism leaves meanings radically underdetermined, even for expressively rich languages. Here philosophers usually appeal to sophisticated results from model theory, but the point is quite trivial. If the only interpretative constraint is that some total theory should come out true (or largely true), one can simply assign arbitrary true propositions to the sentences in the theory. For instance, we can interpret all sentences as meaning that salt dissolves in water. The results from model theory show that radically different and absurd interpretations are still possible if we hold fixed the interpretation of logical constants and the rules of compositional semantics, but Global Descriptivism does not allow us to take any meaning facts as externally fixed.[8]

In 'Putnam's Paradox', Lewis discusses the undetermination problem. To avoid it, he suggests that one might appeal to the objective naturalness of referents. Thus we arrive at the magnetic account quoted in section 1:

8 Analogous problems arise for Global Descriptivism as a meta-semantics for propositional attitudes – a form of "conceptual role semantics" attributed to Lewis in [Stalnaker 2004b] (see fn.9 below). Here the idea would be that agents have a certain set of sentence-like representations in their head, whose content is given by whatever interpretation makes most of them come out true. Again, this ignores many facts that are clearly relevant to the content of an intentional state, such as the conditions under which the state typically occurs, the (non-verbal) behaviour it typically causes, etc. It also makes expressive redundancy a precondition of mental representation, and it radically underdetermines the content of mental states.

> overall eligibility of reference is a matter of degree, making total theory come true is a matter of degree, the two desiderata trade off. The correct, 'intended' interpretations are the ones that strike the best balance. [Lewis 1984: 65f.]

The suggestion is puzzling. It doesn't help with the underdetermination problem. Suppose negative unit charge is a perfectly natural property. Then we can interpret any given set of sentences by assuming that (a) all words refer to negative unit charge, and (b) every sequence of words referring to negative unit charge means that salt dissolves in water. (Part (b) is the compositional semantics.) We have found a simple and maximally natural interpretation that makes the set of sentences true.

Even if we take the compositional rules and the meaning of logical constants as externally given, it is unlikely that the naturalness constraint can rule out deviant interpretations, as [Williams 2007], [Hawthorne 2007] and others have pointed out. Moreover, imposing strong naturalness constraints on reference threatens to miss the actual meaning of words like 'garden', 'green' or 'grue', whose reference is not particularly natural.

What is going on here? Why does Lewis defend magnetised Global Descriptivism in 'Putnam's Paradox'? I think one possibility can be safely ruled out. Global Descriptivism, with or without magnetism, was not Lewis's own view. It flatly contradicts almost everything he wrote elsewhere on language, both before and after 1984 (like [Lewis 1986b: 40–50] and [Lewis 1992]). At most, we could assume that for a very brief period in the early eighties, Lewis gave up the brilliant account of language he had developed in great detail since the mid sixties and replaced it by an utterly ridiculous alternative, without anywhere pointing out this change of mind and only to return to his old view shortly afterwards. This is incredible.

So Lewis defended a position he didn't actually endorse. This was not unusual for him; once he published a paper under the pseudonym of his cat, [LeCatt 1982], in which he argued against one of his own theories. In the case of 'Putnam's Paradox', his paper is a response to Putnam, and Putnam's arguments presuppose something like Global Descriptivism: that the correct interpretation of our words merely has to render our (ideal) total theory true. Lewis himself rejected this approach to meta-semantics. However, he thought that underdetermination arguments similar to Putnam's could be mounted against a simplistic version of his functionalist account of mental content, and that (*via* principles of humanity) considerations of objective naturalness should be used to resolve the underdetermination. We also saw that on Lewis's own view, it is in part due to the greater naturalness of *green* over *grue* (or in any case due to *some* connection between these properties and objective naturalness) that 'emeralds are green' means that emeralds are green rather than that they are grue – although the reason is complicated and indirect. So by his own lights, Lewis's response to Putnam contained at least two important grains of truth.

Fortunately, Lewis is fairly explicit about all this. His paper begins with a couple of

caveats. One is that Lewis will assume for the sake of Putnam's argument that semantic values can be identified with referents. As we saw above, Lewis did not himself believe that. Here is the other caveat:

> 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. [Lewis 1984: 57f.]

He adds a footnote: "For a discussion of the 'relocated' problem and its solution, see the final section of my 'New Work for a Theory of Universals' ".

Looking up the final section of 'New Work for a Theory of Universals', we find – after an abridged version of 'Putnam's Paradox' on pp.45–49 and again the note that the problem arguably rests on a misguided approach to meta-semantics – exactly the story about intentionality that I've told in the previous sections: that mental content is determined by a state's functional role, that the folk psychological input and output conditions do not suffice to rule out all deviant interpretations, that principles of humanity are needed as further constraints, and that some of those princiles involve the distinction between natural and unnatural properties. This, then, is the properly relocated problem and its solution.[9]

9 Robert Stalnaker [2004b] reads the passage just quoted from 'Putnam's Paradox' as indicating that Lewis endorsed magnetised Global Descriptivism as his theory of mental content. This is a very strange interpretation. It not only ignores the attached footnote, it is also completely at odds with Lewis's numerous writings on intentionality both before and after 1984. Stalnaker notes that his interpretation does not square with things Lewis says elsewhere, but takes this to be a problem for Lewis. How could he think that 'Putnam's Paradox', of all places, contains Lewis's final theory of mental content? (Why not consult, say, the section 'Content' in 'Reduction of Mind' [Lewis 1994]?) The answer, I suspect, is that Stalnaker failed to distinguish two notions of *narrow content*. Lewis always maintained that mental content is narrow in the sense that functional duplicates in the same world share their beliefs and desires, no matter whether they live on Earth, on Twin Earth, or in a vat (see e.g. [Lewis 1979a: 142f.], [Lewis 1994: 312–324]). Stalnaker disagrees and holds that the content of our beliefs is in part determined by the objects to which we are actually causally related – H_2O, for example, and not XYZ. He seems to assume that the only way to avoid the consequence that content is in this sense wide is to assume that mental content is determined without any recourse to external causes and effects (see [Stalnaker 1993], [Stalnaker 2004a]). This would indeed only leave room for something like phenomenalism or a mentalese version of Global Descriptivism. But Lewis's mental content is not narrow in the sense of being determined independently of external causes and effects. On the contrary, such causal connections play a big role in Lewis's folk psychological analysis.

6 Against magnetism

At the core of Lewis's philosophy of language is the idea that semantics should be part of a more comprehensive theory of human interactions. The point of semantics is not simply to record or systematise intuitions about meaning, truth and reference, any more than the point of genetics is to systematise people's intuitions about genes and inheritance. Rather, the semantical pairing of sentences with truth-conditions reflects interesting patterns in the social interactions among the members of a linguistic community.

Suppose I know something that I want you to know – say that we are being followed by a tiger. How can I get this across to you? I can produce sounds or other signs. But how can I know that you will take my signs as indicating that we are being followed by a tiger, and not, for example, that dinner is ready? Here it helps to have a convention that links my signs to the presence of tigers. Suppose it common knowledge between you and me that we try to produce those signs only when we are being followed by a tiger, or when we want the addressees to realise that they are being followed by a tiger. Then I can trust that you will understand what I want to achieve when I produce the signs. The signs can serve as an efficient means to get across what I want to get across. In general, for language to serve its communicative function, there must be something like a shared association between sentences and possible states of affairs. For Lewis, this association is the topic of semantics.

The basic idea can be fleshed out in many ways. We do not have to accept Lewis's particular analysis of linguistic convention, with its focus on rationality, common knowledge and shared interests. (Stripped-down versions of Lewis's account have been fruitful in theories of signaling between non-human animals and even bacteria; see [Skyrms 2010].) We could also take into account neuropsychological information about the internal processes that take place in actual language production and reception, thereby sacrificing a certain level of generality but improving the integration with empirical theories of human cognition. With respect to the basic shape of semantic theories, we could let them pair expressions with dynamical rules for updating states of information, rather than pair sentences with truth-conditions. More obviously, we do not have to follow Lewis's functionalist account of mental content; we do not even have to assume that mental content is prior to linguistic content. Whichever way we go, we can preserve the idea that semantics serves an explanatory purpose in a more comprehensive theory of human (or non-human) interactions.

And here lies the problem with magnetism. Suppose a sentence S is systematically used to communicate that a certain state of affairs p obtains: speakers try to utter S only if p obtains, hearers take utterances of S as evidence for p, and so on. If magnetism is true, then it is a live possibility that, unbeknownst to everyone, S actually means not p but q, because q concerns objectively more natural properties. The challenge for

the magnetist is to explain what this conception of meaning should be good for. Which aspects of the practice of communication and coordination are better explained by a semantics that pairs S with q rather than p?

To be sure, there are good reasons to introduce a notion of "literal" meaning that does not simply capture the information typically expressed or conveyed by utterances of a sentence. One such reason is that sentences can be embedded in bigger sentences, and we might want the truth-conditions of bigger sentences to be determined by the meanings and syntactical arrangement of their parts. For this purpose, meanings have to go beyond conventional truth-conditions. Another reason is that there is usually a presumption that speakers try to make statements that are helpful, reasonable and relevant to the conversational context; it then follows that a sentence that means one thing can be systematically used to convey something else (typically something stronger), because the utterance would not be helpful or relevant unless the speaker believed the other thing. Nothing like this can be said in defense of the magnetist's distinction between the alleged meaning q and the information p that is actually conveyed.

Consider a hypothetical example. Imagine a community of language users that only eat root vegetables and a rare type of mushroom. They have a word 'food' that plays a role similar to 'food' in English, which they apply to root vegetables as well as the mushroom. But the root vegetables by themselves form a much more natural category than the root vegetables together with the mushroom. According to reference magnetism, the community's word 'food' might therefore pick out only root vegetables, since that is the more natural referent. How does this help explain or systematise the community's linguistic practice? What is the explanatory advantage of the magnetic interpretation?

People sometimes distinguish "trumping" from "tie-breaking" forms of magnetism. What I call magnetism is trumping magnetism. Tie-breaking magnetism would be the view that objective naturalness only enters the picture to resolve indeterminacies of meaning: if different assignments of meaning equally fit all aspects of use, then the true meanings are the most natural candidates. I will not dwell on this proposal, since it would not have the consequences endorsed by Weatherson and Sider, but I think the present objection also applies to tie-breaking magnetism. Suppose our attitudes and conventions do not settle exactly what shades of colour fall under 'red' in a given context, or exactly which towns and villages belong to what we call 'the outback'. Suppose, moreover, that some of the candidates are objectively more natural than others. Should we conclude that 'red' and 'the outback' determinately pick out these candidates? I do not see the advantages of such a proposal. What does it explain that could not be explained by the other possible choices? In fact, a systematic theory of our linguistic practice should arguably not make any choice among the candidate meanings, but instead reflect the extent of indeterminacy and vagueness. Among other things, this has a better chance of explaining the use of modifiers like 'it is indeterminate whether', the pragmatic shiftiness

of borderlines, and our reluctant and sometimes contradictory judgments about cases on the border.

7 Illusions of magnetism

There is a real and widespread phenomenon that might at first glance be taken to support magnetism; a well-known instance concerns fish. Before the discovery that whales are mammals, the word 'fish' was commonly used for both fish and whales. Nevertheless, it is often intuited that 'fish' did not mean *either fish or whale*. Rather, 'fish' always meant *fish*, and our ancestors were simply mistaken about the whales.

I do not know whether this is historically true, but it is certainly conceivable. Perhaps our ancestors used 'fish' with the convention that it is to pick out a biologically homogeneous class of objects. We can imagine that the word was introduced by a stipulation to the effect that it picks out the biologically most natural kind including such-and-such exemplars (here we point at some carp and herrings) and excluding such-and-such others (crabs, snails, elephants). Since the fish form a much more natural biological kind than the fish together with the whales, this would mean that whales do not fall under the predicate 'fish', irrespective of whether anyone is aware of that fact.

In reality, the word 'fish' was of course not introduced by explicit stipulation, but it may well have been used as if governed by that stipulation. To adjudicate the question, we would have to ask, for example, whether our ancestors, when they claimed that whales are fish, expressed a false belief that whales belong to the same biological kind as carp and herrings (but not snails and elephants), or whether they rather expressed the true belief that whales share certain superficial traits with carp and herrings. Or we could ask how they would have responded to the discovery that whales are biologically very different from carp and herrings and more like hippos and cows: would they have treated 'whales are not fish' as true? I suspect their usage may not have been fully determinate in these respects. But it certainly makes sense to have words tracking unknown natural boundaries in the world.

Many terms in science have this pattern of use. In physics, for example, 'temperature' is used to pick out a physically significant quantity that stands in certain relations to other quantities and to everyday experience. As Norton [forthcoming] points out, this creates a kind of error tolerance for theories of temperature. In classical thermodynamics, temperature is assigned a role that is incompatible with its role in present-day statistical mechanics. Nevertheless, it is plausible that 'temperature' has not changed its reference. There is no physically significant quantity that satisfies the thermodynamic statements about temperature. There certainly are functions from objects to numbers that do satisfy the equations, but these functions do not partition the world into physically significant classes and therefore do not qualify as possible referents of 'temperature'. There is one, and

only one, natural quantity that comes close to playing the thermodynamic temperature role, namely temperature. This is enough for 'temperature' in thermodynamics to denote that quantity.

These observations show that the reference of words like 'fish' and 'temperature' can depend on objective joints in nature. Some candidate referents are more natural, and thereby more eligible than others. Moreover, naturalness can "trump" other aspects of use: even if everyone says that whales are fish, or that temperature satisfies such-and-such equations, they can be wrong because there is a highly natural referent in the vicinity for which these claims are false.

But this is not magnetism. The reason why naturalness here plays a role in determining reference is that *this is how we use the words.* Nothing forces us to use words like 'fish' and 'temperature' with the convention that they pick out reasonably natural properties. Moreover, there is nothing special about naturalness here. All kinds of features can enter into the truth conditions associated with sentences, and thereby (derivatively) into the application conditions of predicates. Our linguistic practice makes 'fish' pick out a biologically natural class of things, but also a class of things whose members typically have fins and live in the water. For other terms, objective naturalness is irrelevant. We use colour terms to group things by how they look to us, no matter if the grouping is physically gerrymandered. Similarly, the things we call toys or tools or teddy bears are not meant to be unified by physical, chemical or biological similarities. Scientific discoveries about carp could reveal that carp are not fish, but scientific investigations into shovels could never reveal that shovels are not tools. Even for words where naturalness enters into the application conditions, it is not always the same kind of naturalness. A class of objects unified by biological similarities need not be unified by more basic physical or chemical similarities, due to the multiple realisability of biological properties.

Crucially, the hypothesis that various sorts of naturalness sometimes play a role in the conventional truth-conditions of sentences does not introduce any secrets into semantics. If 'fish' is used to pick out a biological kind, then this is common knowledge in the linguistic community. Competent speakers know that 'carp are fish' could turn out to be false, and they know what sorts of discoveries would support its falsehood. When they utter sentences involving 'fish', they communicate information about a biological kind. It's not that one thing is communicated and another thing is the meaning. This also means that we do not get the striking consequences suggested by Weatherson and Sider. Competent speakers can be systematically mistaken in their application of a predicate, if they mistakenly believe that the relevant objects satisfy the conditions associated with the predicate. It is much harder for them to be systematically mistaken about those conditions themselves. There is no reason to believe that we are systematically mistaken about the conditions under which something counts as 'knowledge', or under

which properties count as 'instantiated'.[10]

8 Conclusions

I have argued that although Lewis defends the magnetic conception of meaning in 'Putnam's Paradox' and pp.45–49 of 'New Work for a Theory of Universals', it is clear both from the context of these passages and his other writings that he did not endorse that view. In Lewis's own account of language, objective naturalness plays essentially no role at all. Lewis did see a role for naturalness in his account of mental content, namely to unify and generalise certain constraints on the distribution of agents' probabilities and utilities. But even here the picture would not change dramatically if we left those constraints un-unified and returned to the pre-1983 account.

Exegetical matters aside, I have argued that the magnetic conception should be rejected, because it drives a wedge between meaning and use. A semantics for English or German should fit into a more comprehensive account of how language use helps members of these communities to communicate and coordinate. If a certain sentence is conventionally used to communicate that such-and-such conditions obtain, it should be possible to recover these conditions from the semantic value assigned to the sentence. On the magnetic conception, this can easily fail: the true meaning of the sentence could be something else, merely because the alternative is objectively more natural. Friends of magnetism need to tell us the explanatory advantage of magnetised semantics. I cannot see what it might be.

My target has been magnetism in meta-semantics, as assumed e.g. in [Weatherson 2003] and [Sider 2009]. Other forms of magnetism are not affected by my criticism. I have agreed that, as a matter of fact, the meanings of some words make them trace natural boundaries in the world. I have also raised no general objections to the use of naturalness constraints in the interpretation of mental states. The link between meaning and use only requires that semantic values should not be sensitive to facts beyond the regularities in linguistic behaviour and the psychology of speakers and hearers. Doesn't this leave room for the possibility that due to some naturalness constraint on the content of attitudes, 'knowledge' really does mean justified true belief? Perhaps on the best interpretation of our mental states, we actually *believe* that the word 'knowledge' applies in Gettier cases, and so we *intend* to use the word when describing such cases, and *expect* others to do the same? The proposal is obviously crazy; besides, it gives rise to a problem that should look familiar: since these alleged beliefs, intentions and expectations do not fit our actual behaviour nor our disposition, what explanatory purpose could be served by this assignment of content?

10 I am not the first to suggest that the application conditions of words sometimes involve objective naturalness, see e.g. [Lewis 1994: 313] and [Jackson 1998a: 95].

References

Tyler Burge [1973]: "Reference and Proper Names". *Journal of Philosophy*, 70(14): 425–439

Ralf Busse [2009]: "Properties in Nature. A Nominalist Account of Fundamental Properties". Habilitationsschrift, University of Regensburg, Germany

Donald Davidson [2004]: *Problems of Rationality*. Oxford: Clarendon Press

Richard E. Grandy [1973]: "Reference, Meaning, and Belief". *Journal of Philosophy*, 70(14): 439–452

Paul Grice [1957]: "Meaning". *The Philosophical Review*, 66(3): 377–388

John Hawthorne [2007]: "Craziness and Metasemantics". *The Philosophical Review*, 116(3): 427–440

Frank Jackson [1998a]: *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press

— [1998b]: "Reference and Description Revisited". *Philosophical Perspectives*, 12: 201–218

Richard Jeffrey [1965]: *The Logic of Decision*. New York: McGraw-Hill

James Joyce [1999]: *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press

Bruce LeCatt [1982]: "Censored Vision". *Australasian Journal of Philosophy*, 60(2): 158–162

David Lewis [1966]: "An Argument for the Identity Theory". *Journal of Philosophy*, 63: 17–25. Reprinted with extensions in [Lewis 1983c]

— [1969]: *Convention: A Philosophical Study*. Cambridge (Mass.): Harvard University Press

— [1970a]: "General Semantics". *Synthese*, 22: 18–67. Reprinted in [Lewis 1983c]

— [1970b]: "How to Define Theoretical Terms". *Journal of Philosophy*, 67: 427–446. Reprinted in [Lewis 1983c]

— [1974a]: "Radical Interpretation". *Synthese*, 23: 331–344. Reprinted in [Lewis 1983c]

— [1974b]: "'Tensions". In Milton K. Munitz and Peter K. Unger (Eds.) *Semantics and Philosophy,* New York: New York University Press. Reprinted in [Lewis 1983c]

— [1975]: “Languages and Language”. In Keith Gunderson (Ed.) *Language, Mind, and Knowledge,* vol VII of *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press, 3–35. Reprinted in [Lewis 1983c]

— [1979a]: “Attitudes *De Dicto* and *De Se*”. *The Philosophical Review*, 88: 513–543. Reprinted in [Lewis 1983c]

— [1979b]: “Scorekeeping in a Language Game”. *Journal of Philosophical Logic*, 8: 339–359. Reprinted in [Lewis 1983c]

— [1980a]: “Index, Context, and Content”. In S. Kanger und S. Ohmann (Hg.), *Philosophy and Grammar*, Dordrecht: Reidel. Reprinted in [Lewis 1998a]

— [1980b]: “Mad Pain and Martian Pain”. In Ned Block (Hg.), *Readings in the Philosophy of Psychology* Bd.1, Cambridge (Mass.): Harvard University Press, 216–222. Reprinted in [Lewis 1983c]

— [1980c]: “Veridical Hallucination and Prosthetic Vision”. *Australasian Journal of Philosophy*, 58: 239–249. Reprinted in [Lewis 1986c]

— [1981]: “Causal Decision Theory”. *Australasian Journal of Philosophy*, 59: 5–30. Reprinted in [Lewis 1986c]

— [1983a]: “Individuation by Acquaintance and by Stipulation”. *The Philosophical Review*, 92: 3–32. Reprinted in [Lewis 1999]

— [1983b]: “New Work for a Theory of Universals”. *Australasian Journal of Philosophy*, 61: 343–377. Reprinted in [Lewis 1999]

— [1983c]: *Philosophical Papers I*. New York, Oxford: Oxford University Press

— [1984]: “Putnam’s Paradox”. *Australasian Journal of Philosophy*, 61: 343–377. Reprinted in [Lewis 1999]

— [1986a]: “Against Structural Universals”. *Australasian Journal of Philosophy*, 64: 25–46. Reprinted in [Lewis 1999]

— [1986b]: *On the Plurality of Worlds*. Malden (Mass.): Blackwell

— [1986c]: *Philosophical Papers II*. New York, Oxford: Oxford University Press

— [1986d]: “Probabilities of Conditionals and Conditional Probabilities II”. *The Philosophical Review*, 95: 581–589. Reprinted in [Lewis 1998a]

— [1989]: “Dispositional Theories of Value”. *Proceedings of the Aristotelian Society*, Suppl. Vol. 63: 113-137. Reprinted in [Lewis 2000]

— [1992]: "Meaning without Use: Reply to Hawthorne". *Australasian Journal of Philosophy*, 70: 106–110. Reprinted in [Lewis 2000]

— [1993]: "Many, But Almost One". In J. Bacon, K. Campbell and L. Reinhardt (Eds.) *Ontology, Causality and Mind: Essays in Honour of D.M. Armstrong,* Cambridge: Cambridge University Press, 23–38. Reprinted in [Lewis 1999]

— [1994]: "Reduction of Mind". In Samuel Guttenplan (Hg.), *A Companion to the Philosophy of Mind*, Oxford: Blackwell, 412–431. Reprinted in [Lewis 1999]

— [1996]: "Elusive Knowledge". *Australasian Journal of Philosophy*, 74: 549–567. Reprinted in [Lewis 1999]

— [1997]: "Naming the Colours". *Australasian Journal of Philosophy*, 75: 325–342. Reprinted in [Lewis 1999]

— [1998a]: *Papers in Philosophical Logic*. Cambridge: Cambridge University Press

— [1998b]: "A World of Truthmakers?" *Times Literary Supplement*, 4950: 30. Reprinted in [Lewis 1999]

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

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

— [2001]: "Redefining 'Intrinsic' ". *Philosophy and Phenomenological Research*, 63: 381–398

— [2004]: "How Many Lives has Schrödinger's Cat? The Jack Smart Lecture, Canberra, 27 June 2001". *Australasian Journal of Philosophy*, 82: 3–22

— [2009]: "Ramseyan Humility". In D. Braddon-Mitchell and R. Nola (Eds.) *Conceptual Analysis and Philosophical Naturalism,* Cambridge (Mass.): MIT Press, 203–222

Richard Montague [1973]: "The Proper Treatment of Quantification in Ordinary English". In J. Hintikka and J. Moravcsik (Eds.) *Approached to Natural Language,* Dordrecht: Reidel, 221–242

John D. Norton [forthcoming]: "Dense and Sparse Meaning Spaces". In R.M. Burian and A.Gotthelf (Eds.) *Concepts, Induction, and the Growth of Scientific Knowledge,*

Leonard Savage [1954]: *The Foundations of Statistics*. New York. Wiley

Theodore Sider [2004]: "Symposium on *Four-Dimensionalism*". *Philosophy and Phenomenological Research*, 68: 642–647

— [2009]: "Ontological Realism". In D. Chalmers, D. Manley and R. Wasserman (Eds.) *Metametaphysics,* Oxford: Oxford University Press, 384–423

— [2011]: *Writing the Book of the World*. Oxford: Oxford University Press

Brian Skyrms [2010]: *Signals*. Oxford: Oxford University Press

Robert Stalnaker [1993]: "Twin Earth Revisited". *Proceedings of the Aristotelian Society*, 93: 297–311

— [2004a]: "Assertion Revisited: On the Interpretation of Two-Dimensional Modal Semantics". *Philosophical Studies*, 118: 299–322

— [2004b]: "Lewis on Intentionality". *Australasian Journal of Philosophy*, 82: 199–212

Brian Weatherson [2003]: "What Good are Counterexamples?" *Philosophical Studies*, 115: 1–31

— [Manuscript]: "The Role of Naturalness in Lewis's Theory of Meaning".

J. Robert G. Williams [2007]: "Eligibility and Inscrutability". *The Philosophical Review*, 116: 361–399