Can we explain linguistic representation in terms of perceptual representation?

Jeff Speaks
October 26, 2011

1. The structure of mentalist theories of meaning

The most well-worked out attempts to explain the meanings of expressions in public languages are mentalist theories of meaning: theories that attempt to explain linguistic representation in terms of mental representation.

Such theories can be thought of as part of a two-part mentalist theory of intentionality. In **stage 1**, we identify a class of *meaning-constituting mental states* in terms of which meaning is to be explained, and say what relation between these mental states and a linguistic expression is necessary and sufficient to endow that expression with certain meaning. The usual candidates for meaning-constituting mental states are intentions (as in Grice) and beliefs (as in Lewis). In **stage 2**, we try to explain the contents of the meaning-constituting mental states — preferably in terms which don't involve any representational facts.

In this paper I want to do a few things: first, raise some difficulties for these usual candidates; second, explore the possibility that perceptual states might be better candidates for meaning-constituting mental states; and, third, say why this choice would make the execution of stage 2 of a mentalist theory of intentionality more difficult.

2. Problems with the usual candidates

Attempts to carry out stage 1 of the reduction of meaning — either in the manner of Grice or in the manner of Lewis — have given rise to a small industry of counterexamples. However, counterexamples often motivate revision of the analysis rather than abandonment of the research program; so it would be interesting to see whether

more general arguments can be given against these views. That's what I will try to do in this section.

Before moving on to these arguments, though, it's worth being a bit clearer on the basis for our taxonomy of mentalist theories of meaning. Above I suggested that we could distinguish between the Gricean and the Lewisian analyses of meaning by their choice of meaning-constituting mental states: the Gricean analyzes meaning in terms of communicative intentions, and the Lewisian analyzes meaning in terms of beliefs. But this is, in one respect, over-simple, since (at least some versions of) both the Gricean and the Lewisian analysis make use of *both* intentions and beliefs. While discussion of Grice often focuses on his analysis of speaker-meaning in terms of intentions, the Gricean must also provide an analysis of expression meaning in terms of speaker-meaning, and it is natural to analyze the meanings of expressions in a community in terms of mutual *belief* about what members of that community mean, or would mean, by various uses of those expressions.[1] Just so, the Lewisian explains the meanings of a sentence in terms of the content of a belief with which it is conventionally associated — but in explaining what it is for a convention to be operative in a group, one might well appeal to *intentions* of members of that group to, for example, utter only sentences that they believe to be true.[2]

But even if the Gricean and the Lewisian both make use of both intentions and beliefs, they give them importantly different roles. Mentalist theories typically give one class of mental states a certain central role in the theory: the role of explaining the meaning of expressions *directly*, by giving the expression in question *their own content*. Communicative intentions play this role for Grice — sentences end up with *the same content as* the communicative intentions with which they are associated (even if explaining what "associated" means here involves, for some neo-Griceans, appealing to the beliefs or other mental states of the relevant speakers). And, for Lewis, sentences have *the same content as* the beliefs with which they are conventionally associated (even if explaining what it is for a convention to be operative in a group involves thinking about the intentions of members of that group). More generally, we can distinguish two different roles that mental states of a certain type can play in a mentalist theory. First, they can be among the states which confer their own content on the sentences or other expressions to which they bear the right relations — these are what I call the *meaning-constituting mental states*. Second, they can be part of the story about what this "right relation" between sentences and the meaning-constituting mental states is. I'll call the mental states which figure in the explication of this "right relation" *linking mental* states. Though the Gricean and the Lewisian make use of many of the same mental states in their theories, they differ in which states they accord the fundamental, meaning-constituting role of conferring their own content on the expressions of the language.

---

[1] This is the approach of (among others) Schiffer (1972); Grice did not take this route. See his discussion of resultant & basic procedures in Grice (1968).

[2] Lewis doesn't explicitly make use of intentions, though he does rely on facts about what speakers try to do, as well as facts about their (intentionally) uttering, and avoiding uttering, certain sentences. See e.g. Lewis (1975), p. 167.

2.1. Meaning and intention

With that preliminary aside, let's consider first the Gricean analysis. As is well known, Grice's attempt to execute stage 1 of a mentalist theory of intentionality itself has two stages: an analysis of sentence meaning in terms of speaker-meaning, and an analysis of speaker-meaning in terms of communicative intentions. There are many different versions of the latter sort of analysis in the literature; a representative example is the following:

> [G] *A* means *p* by uttering *S* iff *A* intends in uttering *S* that:
> (1) his audience come to believe *p*,
> (2) his audience recognize this intention, &
> (3) (1) occur on the basis of (2)[3]

[G] has a feature common to most Gricean analyses: it holds that it is a necessary condition on a speaker meaning *p* by an utterance that the speaker intend that the audience form a certain belief (in this case, the belief whose content is the proposition meant by the speaker).

A problem with this is that sentences of the following form can be seriously and truly uttered:

> [K] *S*, but I know that you will not believe that *S*.

(Imagine an unfaithful spouse saying: "I love you, though I know you won't believe that.") Suppose that *S*, in the context, expresses the proposition *p*. We then aim to show that (i) the speaker means *p* by his utterance, and (ii) the speaker does not intend that his audience believe *p* — which together entail the falsity of [G].

> *Argument for (i).* The speaker was speaking seriously, so in this sort of case the speaker will typically mean by his utterance the conjunctive proposition the sentence uttered semantically expresses. But speaker-meaning (like assertion, communication, belief, and many other related propositional attitudes) distributes over conjunction; hence it follows that the speaker means *p* (the proposition expressed by the first conjunct of the sentence uttered) by his utterance.

> *Argument for (ii).* By hypothesis the sentence was true, so the speaker knows that his audience will not believe *p*. But there are epistemic constraints on intention and, in particular, it is a necessary truth that we cannot intend to bring about effects which we know our actions will not bring about. Hence it follows that the speaker does not intend that his audience believe *p*.

---

[3] See, for example, Grice (1957).

The interest of this argument is not so much its force as an argument against [G] in particular as its generality.

To see the generality of this line of objection against various versions of the Gricean analysis, consider the view — developed as a successor to [G] in order to handle cases like an oral examination, in which the speaker means something by an utterance despite knowing that it is already believed by her audience — that revises clause (1) of [G] to require that a speaker intend that her audience believe *that she believes that p*.[4] This is open to just the same argument as that just given against [G], by varying the problematic sentence to

[K*] *S*, but I know that you will not believe that I believe that *S*.

Again, the example of the unfaithful spouse — this time expecting to be regarded as dishonest rather than merely self-deceived — serves to make the point.

And, in general, if we suppose the Gricean modifies his analysis so that clause (1) requires not the intention that one's audience form the belief that *S*, but rather that they stand in some other propositional attitude relation *R* to some other proposition *q,* the present argument can be adapted to refute this new analysis, if we can formulate some suitable sentence

S, but I know that you will not bear *R* to *q*.

which, like the one above, can be uttered seriously and truly.

Objection 1: intending to bring about x is compatible with knowing that you will not bring x about. Two replies: (i) surely there are some epistemic constraints on intention and, whatever they are, the above sentence can be reformulated accordingly, e.g. "*S*, though I am completely subjectively certain that you will not believe that *S*." (ii) the Gricean who denies that there are any epistemic constraints on intention at all is left without an explanation of a phenomenon involving speaker-meaning to which Wittgenstein drew attention in the *Investigations*. Wittgenstein tried to imagine someone saying "a b c d" and meaning by this that the weather is fine.[5] Could one do this? Could one simply, in a conversation, decide to mean that the weather is fine by an utterance of "a b c d"? Or consider another example from Wittgenstein:

> "Make the following experiment: *say* "It's cold here" and *mean* "It's warm here." Can you do it?" (§510)

[4] Grice himself came to think that speaker-meaning should be partly analyzed in terms of intentions that an audience come to believe something about the speaker's beliefs; see Grice (1969).

[5] See Wittgenstein (1953), §508.

I'm inclined to agree with Wittgenstein that this is not possible (absent, of course, special stage setting in which we've decided to speak in code). The case has, to me, the same sort of feel to it as the attempt to believe something at will.

But what could explain the impossibility here, if speaker-meaning is, as the Gricean claims, just a matter of having certain communicative intentions? The only plausible answer seems to be: it is impossible to form the relevant intentions in this sort of case, because one knows that the utterance won't bring about the intended effect. But this relies on there being some epistemic constraint on intention — which is just the premise which the above form of argument requires.

Objection 2: the objection is not general, since there are some analyses for which no suitable sentence of the above form can be found; for example, analyses which let the relevant propositional attitude be one which no speaker could doubt his utterance will cause. Probably there are such analyses, but the very aspects of them which make them immune to this argument make them implausible candidates for the analysis of speaker-meaning. Take as an example a suggestion from Neale (1992): perhaps meaning p requires only the intention that one's audience entertain the thought that p. There are two related problems here. First, it requires dropping clause (3) of [G], since one's audience will in the standard case entertain the thought regardless of whether they recognize that the speaker intends them to, and the speaker might well know this; and, second, once we drop clause (3), the conditions on speaker-meaning become far too weak. I can intend that my audience entertain the thought that I am speaking (or saying certain words), and intend that they recognize my intention, without meaning by my utterance that I am speaking.

2.2. Meaning and belief

Let's turn now to the idea that the meaning-constituting mental states are not communicative intentions, but the beliefs of language users.

Such an analysis will be an instance of the schema:

> [L] S means that p in a given community iff for every (or almost every, or most, or ...) member(s) of that community, S bears R to the belief that p.

The trick is then to specify the relation R. A simple and natural candidate is the relation corresponding to the open sentence

> A seriously utters S only if A believes that p

which gives us the analysis

> [L1] S means that p in a given community iff every (or almost every) member of the community seriously utters S only if she believes that p

More sophisticated versions of [L] — including the version defended by Lewis — are available. But I want to begin with a simple problem for [L1] which, I think, is also a problem for more sophisticated versions of the analysis of meaning in terms of belief.

This problem arises from a fundamental distinction between facts about meaning and facts about belief: facts about belief are *clumpy*, in the sense that there are distinct beliefs such that, necessarily, anyone who has one belief also has the other. One example comes from the distribution of belief over conjunction: it is a necessary truth that if someone believes a conjunctive proposition, they also believe both of the conjuncts. By contrast, facts about the meanings of sentences are not clumpy in this way; it is never a necessary truth, for a pair of propositions, p, q, that any sentence which semantically expresses one must also semantically express the other. But putting these two facts together with [L1] entails a contradiction. By [L1] plus the distribution of belief over conjunction, it follows that every conjunctive sentence must semantically express, not just the relevant conjunctive proposition, but also the propositions expressed by its conjuncts — which, given the non-clumpiness of the facts about meaning, is false.

While this example uses beliefs which are *necessarily* connected, we needn't rely on such strong connections between beliefs; all that is needed to provide a counterexample to [L1] are beliefs which are such that everyone in the community who has the first also has the second.[6] Suppose, for example, that in a certain community fear of dogs is universal, and that everyone who believes that there is a dog before her also believes that there is a dangerous beast before her; surely this is not sufficient for any sentence which means that there is a dog before me to also mean that there is a dangerous beast before me.[7]

It is useful to contrast this objection with one which John Hawthorne pressed against the instance of [L] defended by David Lewis. Hawthorne pointed out, in effect, that natural languages like English contain many sentences too long for any speaker of the language to utter.[8] Consider any such sentence S, and two propositions, p and q, which are candidates for being the semantic content of this sentence. It doesn't seem as though we can decide between these two propositions by asking: which proposition is such that no speaker would utter S without believing it? — since, after all, no speaker would utter

---

[6] Even this is a bit too strong, since one might think that the relation R only has to hold between almost all, or most, or many agents and the relevant sentences and propositions; in that case the relevant beliefs would only have to be coinstantiated in almost all, or most, or many members of the community.

[7] This shows that a natural response to the example of the distribution of belief over conjunction — namely, that in cases in which multiple propositions satisfy the conditions for being the meaning of S, we should choose the logically strongest as its meaning — won't work. And really this wouldn't work even if we restricted ourselves to examples of pairs of propositions which are such that, necessarily, anyone who believes one believes the other. Consider, for example, the propositions expressed by a pair of sentences S and ⌜Actually, S.⌝ It is plausibly a necessary truth that anyone who believes the proposition expressed by the latter will also believe the proposition expressed by the former — but where S expresses a contingent truth, the former will be logically stronger than the latter, which would lead to ⌜Actually, S⌝ being assigned the proposition expressed by S as its semantic content.

[8] Details are a bit different here, since I'm not considering the full Lewisian analysis. But the point is, for our purposes, the same.

S, period. The question then arises: how can an account of Lewis' sort give an interpretation of the too-large-to-understand fragment of English in terms of the "manageable" fragment of English?[9]

The objection which I'm pressing against [L1] is in one respect more basic: it says that the correlations between beliefs and utterances of sentences don't suffice to determine an interpretation of even the manageable fragment of English, since sentences might be (in the above sense) paired with more than one belief, and known to be so, without being ambiguous.

Let's now consider some ways in which [L1] might be complicated, and ask whether they avoid this objection from the clumpiness of facts about belief and the non-clumpiness of the facts abut meaning. First, one might suggest that meaning requires more than a correlation between sentences uses and beliefs; this correlation must also be mutually known by members of the relevant community:

> [L2] *S* means that *p* in a given community iff
> (i) every (or almost every) member of the community seriously utters *S* only if she believes that *p*, and
> (ii) (i) is mutually known by the members of the community

This is open to just the same sort of counterexample as [L1] — provided we stipulate, as is surely possible, that the facts about the clumpiness of belief used against [L1] are mutually known to obtain.

To block this sort of counterexample, one might — following Lewis (1975) — add another clause to the analysis:[10]

> [L3] *S* means that *p* in a given community iff
> (i) every (or almost every) member of the community seriously utters *S* only if she believes that *p*,
> (ii) (i) is mutually known by the members of the community, and
> (iii) (i) is true because of (ii).

The idea here is that meaning requires more than just mutual knowledge of certain correlations between language use and belief — it requires that these correlations obtain because of the mutual knowledge. Speakers utter certain sentences only when they have certain beliefs because they know that other members of the community follow this practice and expect them to do so as well.

---

[9] This is what leads Lewis and Hawthorne into the topic of whether there is a way of distinguishing "bent" from "straight" grammars; see Hawthorne (1990), Lewis (1992), and O'Leary-Hawthorne (1993).

[10] [L3] is not the full Lewisian analysis — for one thing, it ignores the requirement that there be an alternative regularity which is such that its being generally conformed to would give speakers reason to conform to it. But this is irrelevant to the argument at hand and, as Burge (1975) argues, makes the conditions on meaning too strong.

But it's just not obvious that this extra condition really helps. Consider again the example above, in which everyone who believes that there is a dog in the vicinity also believes that there's a dangerous beast in the vicinity. Suppose that this is mutually known, and hence that it is mutually known *both* that every member of the community utters "Lo, a dog!" (or whatever) only if they believe that there's a dog in the vicinity and that they utter this sentence only if there is a dangerous beast in the vicinity. Suppose now that members of the community sustain this regularity because of their mutual knowledge of *each* of these facts — there's surely nothing to stop this being the case, and its being the case would not suffice for these sentences to be synonymous.

Objection 1: in the above case, everyone who believes that there is a dog before her also believes that there is a dangerous beast before her, but not vice versa; hence we can block counterexamples of this sort by changing, for example, analysis (L1) to say that:

> every member of the community is such that: they seriously utter S *if and only if* they believe that *p*.

Reply 1: this makes the analysis wildly implausible, since people can have unexpressed beliefs.

Reply 2: counterexamples of the above sort can be varied so that the relevant distinct beliefs covary — suppose, for instance, that the community described above believes that only dogs are dangerous beasts. Or consider the proposition that there is a dog in the room and the proposition that ascribes truth to that proposition.

Objection 2: Maybe we can get around these problems by considering a more sophisticated version of the analysis. Our intuition that in the above case "dog" and "dangerous beast" are non-synonymous comes from the facts that the latter is semantically complex, and that "dangerous" has other, non-dog-related, occurrences in the language. But doesn't this just indicate that we should move in the direction of a more complex theory, which maps propositions onto sentences on the basis of a "best fit" between serious utterances of those sentences and beliefs held by speakers, where "best fit" will involve some constraint to do with common propositional constituents being assigned to sentences with words in common?[11]

Reply 1: This modification of the analysis makes substantial assumptions about propositions — namely, that they in some sense have constituents which correspond to elements in the sentences which express them. So they can't be, on this sort of view, sets of possible worlds, situations, etc. This is an assumption

[11] One complication here: we need to modify the added constraint to take account of ambiguity and indexicality, since there won't be, for example, a common propositional constituent corresponding to utterances of all sentences which involve "I" or "bank."

which at least some proponents of this sort of analysis — like Lewis — would not have been willing to make.[12]

Reply 2: It's not clear that the above problem depends on the use of semantically complex expressions. Consider a community which is such that, for some pair of distinct but simple properties F and G, every (or almost every, or ...) member of the community believes that something instantiates F iff they believe that it instantiates G, and this fact is mutually known. This fact about the community should not be sufficient for any pair of simple predicates in their language which, respectively, designate F and G to be synonymous. But, on the present analysis, it would be.

Here's an example: some medievals believed that rationality and risibility were both essential properties of human beings. Suppose they also came to believe that nothing else could have either of these qualities. Then they would believe that, as a matter of necessity, x is rational iff x has risibility. But this wouldn't, I think, make "risible" and "rational" synonyms in their language.

3. Sketch of a mentalist theory of meaning

So much for criticisms of the usual candidates. In this section I want to sketch some more positive thoughts about how we might construct a mentalist theory of meaning which avoids the problems with intention- and belief-based theories. I make no pretense to be giving a theory which approaches the detail in which Grice and Schiffer, among others, have developed intention-based theories, or in which Lewis and others developed a belief-based theory. It's not that I don't think that such a detailed theory is what we should be aiming for — just that I don't know how to give one.

I think that we can make some progress in seeing what form a mentalist theory should take by beginning with a very general question: Why are philosophers attracted to mentalism in the first place?

3.1. Two kinds of supervenience

One important reason is that they find plausible the idea that the meanings of expressions in natural languages (asymmetrically) supervene on the propositional attitudes of users of those languages. One might wonder what explains this necessary connection between mental states and meaning; and a natural and plausible answer is that this necessary connection is to be explained by an analysis of the nature of meaning in terms of certain mental states: there can be no change in semantic properties without a corresponding change in the contents of the mental states of someone or other because

[12] It's true that Lewis was willing to think of "structured intensions" as the meanings of sentences. But what the present modification of the analysis requires is not that structured intensions be the meanings of sentences, but that they be the contents of beliefs — and Lewis would not have been willing to say that this structure was a necessary attribute of the contents of beliefs.

what it is for an expression to have a certain meaning just is for users of the expression to be in certain mental states.

However, to get a clearer picture of the relationship between mental and linguistic representation, it's useful to see that there's an interesting sort of supervenience failure in the vicinity as well. We can formulate two supervenience theses which differ in their treatment of time. Speaking informally, let's say that the A-properties *world-historically supervene* on the B-properties iff every two world histories alike with respect to their B-properties are also alike with respect to their A-properties; let's say that the A-properties *world-slice supervene* on the B-properties iff every two world-slices — every two "worlds at a time" — alike in their B-properties are also alike in their A-properties.[13]

The difference between these two sorts of supervenience is brought out nicely by an example from Wittgenstein:

> "Let us imagine a god creating a country instantaneously in the middle of the wilderness, which exists for two minutes and is an exact reproduction of a part of England, with everything that is going on there for two minutes. Just like those in England, the people are pursuing a variety of occupations. Children are in school. Some people are doing mathematics. Now let us contemplate the activity of some human being during these two minutes. One of these people is doing exactly what a mathematician in England is doing, who is just doing a calculation. — Ought we to say that this two-minute man is calculating? Could we for example not imagine a past and a continuation of these two minutes, which would make us call the processes something quite different?" (*Remarks on the Foundations of Mathematics*, IV, §34)

Wittgenstein here is interested in the question of whether we can say that the two-minute mathematician would be performing an act of the same type as the English

---

[13] I'm not sure that the details matter much here, but I'm thinking of this as a kind of strong supervenience. More formally:

> *World history supervenience*
> Let an order-preserving mapping of individuals and times be a mapping of individual/time pairs between worlds w, w^* such that (i) for any individual i and any time t in w, if $<i,t>$ is mapped to an individual/time pair in w^* whose individual is i^*, then every individual/time pair including i in w is mapped to a pair whose individual is i^* in w^* and (ii) for any times *t1*, *t2*, if *t1* in w is earlier than *t2*, then the time in w^* to which *t1* is mapped must be earlier than the time to which *t2* is mapped. Then the A-properties world-historically supervene on the B-properties iff: For any two worlds w and $w*$, every one-to-one, order-preserving mapping of individuals and times in w and $w*$ such that each individual in one world is alike with respect to the A-properties of the individual with which it is paired in the other world is also such that the paired individuals are alike with respect to their B-properties.
>
> *World slice supervenience*
> For any two world/time pairs w, t and $w*$, $t*$, every one-to-one mapping of individuals in w at t and $w*$ at $t*$ such that each individual in one world is alike with respect to the A-properties of the individual with which it is paired in the other world is also such that the paired individuals are alike with respect to their B-properties.

mathematician. His thought is that we can't, and hence that (in our present, very un-Wittgensteinian terms) facts about intentional act-types don't world-slice supervene on, for example, the physical facts.

We can give an analogous example to show that the facts about linguistic meaning don't world-slice supervene on the facts about mental content:

> In 2695, a historian of American sports is going through some old magazines in the archives, and comes across a story about someone named 'Pete Rose.' The story describes his breaking the career hit record, his being banned from baseball for betting on games, ...

The historian then leaves the archives and says to a colleague, 'I read about this guy, Pete Rose. He was an interesting character ...' When the historian utters this sentence, it seems that the name 'Pete Rose' has a reference for him, and that it refers to the same man to whom it refers for me and other English speakers familiar with the name: a certain late twentieth-century professional baseball player. And this can be so even if in 2695 there are only a few historians of American sports, and so that, immediately prior to this historian delving into the archives, no one had any thoughts, beliefs, or other propositional attitudes about Pete Rose.

But it seems that the name must have had a reference, and hence a meaning, prior to its being discovered by our historian. After all, if it did not have a reference in, say, 2694, then the historian would be effectively introducing a new name into the language. And it would then be puzzling how the historian could do this: how could the historian, who might well lack uniquely identifying information about the referent of the name, be in a position introduce a name which genuinely refers to Pete Rose — as the historian's uses of the name seem to?

An initially attractive line of response to this question is to try to dissolve the puzzle by pointing out that there is a trivial sense in which the archives *do* provide uniquely identifying information about Pete Rose. Just from reading the archives, the historian is in a position to know that if Pete Rose existed, then he is the unique referent of that token of "Pete Rose." Given this, one might think that the historian has simply (in the usual mentalist way, whatever that turns out to be) given the name the meaning of the meta-linguistic description, "the referent of this token of 'Pete Rose.'"

However, this sort of meta-linguistic interpretation of the name out of the researcher's mouth seems (as others have noted) to be open to serious objections:

(1) It seems ad hoc; the researcher does not seem to be expressing thoughts about language, or what refers to what, any more than we usually do.
(2) We seem to get the wrong results when we embed the term to be given the meta-linguistic analysis under modal operators. For example: "Pete Rose could have failed to be the referent of this token of 'Pete Rose'" seems true, but comes out false on the meta-linguistic interpretation. One might try to get around this problem by using a standard descriptivist response to Kripke's modal argument,

and saying that the relevant meta-linguistic description is the *rigidified* description "the referent of this token of 'Pete Rose' in @." But this is open to what seems to me to be a decisive objection from Soames (2002), namely that it is possible to entertain in thought the proposition expressed by "Pete Rose is F" out of the mouth of our researcher without being in a position to think about @.[14]

(3) Just as this view is open to a version of Kripke's modal argument, it also seems to be open to a version of the epistemic argument, because the following seems false: "It is a priori that if Pete Rose exists, then a token of the name 'Pete Rose' exists."

If the researcher does not use the name as an abbreviation for a meta-linguistic description of this sort, then it is hard to see how he could have given the name any meaning which would determine as its reference Pete Rose. Hence the name must have had a meaning before being encountered by the researcher. But, if this is right, then it had this meaning at a time at which there were no speakers with beliefs (or other propositional attitudes) about Pete Rose.

But just as we can imagine Wittgenstein's two-minute world as annexed to different futures and pasts, so we can imagine the slice of our historian's life leading up to his discovery of the article as annexed to different histories and pasts. When we consider different ways of doing this, it seems clear that (i) the meaning of the name "Pete Rose" as used by the researcher would vary depending on which history we choose, and (ii) variation of the history in this way need not affect the propositional attitudes of any users of the language in the time slice in question, before the name is discovered. This is enough to show that meaning does not world-slice supervene on the facts about mental content.

It is very easy to generate examples of this sort involving names, and just as easy to give examples using natural kind terms. (Imagine constructing a case parallel to the one sketched above for "passenger pigeon.") But I think that examples of this sort can also be generated for ordinary predicates. Consider, for example, the color words "red" and "green." Suppose that, over time, red-green color blindness becomes universal in human beings and that, over yet more time, the words "red" and "green" drop completely out of use. Then, one day, long after everyone has stopped having any propositional attitudes about redness or greenness, an intrepid researcher discovers a piece of literature in which the terms "red" and "green" are used. A little investigation tells him that these are color words, and that red and green are distinct colors; but, other than that, he has trouble finding much information about them. Trying to piece together clues from his evidence, he says to one of his fellow researchers: "I think that green is more similar to blue than it is to red."

It seems that the thought he self-ascribes is true, which in turn indicates that he attaches the same intensions to "red" and "green" as we do. But, prior to his acquaintance

[14] One might also reply by saying that the description is not rigidified, but always takes wide scope over modal operators. For what seem to me to be convincing criticisms of this view, see Caplan (2005).

with these terms, he had no thoughts involving concepts with these intensions — which makes it hard to see how he could have, via any mentalist mechanism, given these terms the intensions they have when they are uttered by him. (I'm setting aside the possibility of a meta-linguistic interpretation of his use of color terms, for the reasons given above.) This indicates that these terms had these intensions — and hence presumably contents which determined those intensions — prior to their use by our intrepid researcher. And this, in turn, indicates that the meanings of these color words don't world-slice supervene on facts about mental content — since, as above, we can imagine a different "past and continuation" of this scenario which would give these color terms different meanings without changing anything about the content of any mental state of anyone at the time prior to the discovery.

This line of argument is an objection to theories — like the instances of [L] discussed in the last section — which entail that the meanings of all terms should world-slice supervene on the facts about the beliefs or other propositional attitudes of speakers of the language. But it also suggests some positive conclusions about the relationship between meaning and mental content.

3.2. Two kinds of theory of meaning

The failure of world-slice supervenience suggests that the relationship between meaning and mental content is quite different than the relationship between stock examples of supervening and subvening properties. Consider, for example, the supervenience of shapes on a TV screen on colors of the pixels on the screen — there's simply no room for a gap here between world-historical and world-slice supervenience. The same goes for the supervenience of facts about phenomenal character on intrinsic physical properties; it is of course controversial whether such a supervenience thesis is true, but no one thinks that there's a relevant difference between world-historical and world-slice supervenience here.

What does the fact that matters are different with meaning and mental content show about the relationship between the two? To answer this question it's useful to think about another example of world-historical supervenience without world-slice supervenience: the supervenience of the laws in a certain kind of democracy on the actions of legislators in that democracy.

The explanation of *this* pattern of supervenience is that the laws at a time don't supervene on the actions of legislators at that time because, while legislators do have to perform certain actions to put a law in place, they needn't do anything in particular to ensure that that particular law stays in place. Of course, there are some things they could do which would ensure that it would *not* stay in place, but that isn't the same thing.[15]

The fact that meaning exhibits the same pattern of supervenience on mental states — world-historical but not world-slice — makes it tempting to say that a mentalist theory of meaning should have the same form as an analysis of effective laws in terms of the actions of legislators. That is, perhaps this should be divided into a theory of term introduction

[15] There are other examples of world-historical without world-slice supervenience which seem to point in the same direction, like the supervenience of facts about which objects are works of art on facts about intentional actions.

and a theory of meaning maintenance. On this view, we should think of our theory of meaning not as an attempt to identify a mental property F which makes true

> e means X in group G at $t \equiv$ many/most/almost all members of G have mental property F at t

but rather as an attempt to fill out the following schema with explications of the underlined phrases:

> e means X in group G at $t \equiv$ (i) at some time $t^* \leq t$, e was introduced in G with X as its meaning, and (ii) between t^* and t, e maintained its meaning.[16]

This is, of course, a view of the theory of meaning which is very much like the picture of the reference of proper names sketched by Kripke in *Naming and Necessity*. But despite the influence of Kripke's discussion, writers on the foundations of meaning have not pursued theories of meaning with this two-part structure.

3.3. Meaning maintenance & deference

Examples of semantic shift show that we clearly need *some* substantial theory of meaning maintenance: "must" used to express permission rather than obligation;[17] "egregious" was (according to Wikipedia, at least) once a term for things that were quite good; and as Gareth Evans pointed out, "Madagascar" is a corrupt form of a term which used to refer to part of the African mainland, rather than to the largest African island. Trying to give an account of what all cases of this sort have in common is a highly non-trivial task.

It is also, I think, a task which philosophers of language have unjustly neglected. This accusation might seem a bit unfair, since most theories of the foundations of meaning in the literature don't rely on a distinction between term introduction and meaning maintenance, and hence have a principled reason to think that there is no such thing as the theory I accuse them of not spending enough time on.

But even if most theories of meaning don't overtly have this two-part structure, they almost all make use of a distinction which incurs parallel theoretical obligations. (Here and in what follows I'm indebted to Mark Greenberg's important work on these topics.) Theories of meaning typically give conditions which are supposed to be necessary and sufficient for meaning X by a given expression. But as soon as such conditions are stated, it usually becomes obvious that many speakers seem to be using an expression to mean X despite manifestly failing to satisfy those conditions. These are usually, though not always, cases in which speakers are ignorant or mistaken about some central aspect of the term's meaning. To use the standard examples: one uses "elm" with its normal meaning despite being unable to distinguish elms from beeches; one uses a term to mean *arthritis*

---

[16] Here I'm ignoring for simplicity the possibility of meaning changes of an expression in part of a group, and skirting difficult questions about the individuation of the relevant linguistic communities.

[17] See Traugott and Dasher (2001), §4.2.

despite thinking that the term applies to diseases of the thigh rather than of the joints; one uses "Einstein" to refer to Einstein despite thinking that it refers to someone iff he invented the atomic bomb.[18]

The problem is then what to say about these less than ideally informed individuals (who often turn out to make up most of the language-using population), each of whom, is, as it stands, a counterexample to the relevant theory of meaning. The standard thing to say is that these language users are not genuine counterexamples; rather, they get to use the term with its usual meaning by standing in a certain relation of *deference* to language users whose use *does* satisfy the conditions laid down by our theory (the "experts").[19]

With this in mind, let's return to the example of "Madagascar", mentioned above, in which the name came to stand for an island rather than part of the African mainland. Any theory of meaning owes an account of how the term came to be a name for the island. And any theory of meaning which makes use of the mechanism of deference must, therefore, explain why the later users of the name — who used it to stand for the island — were not just deferring in the usual way to the earlier users of the name, who used it to stand for part of the mainland. But giving an account of this distinction — the distinction between deference and meaning change — seems, at least roughly, extensionally equivalent to giving what I am calling a theory of meaning maintenance.

This brings out the fact that there's a close parallel between the two-part structure advocated above, according to which

> Theory of meaning = theory of term introduction + theory of meaning maintenance

and the more standard approach, according to which

> Theory of meaning = theory of expert use + theory of deference

To some extent, the distinction between them might just be terminological. But I think that there are some substantive differences, and a few reasons to prefer the first way of looking at things.

One is that the relation of deference — in which non-expert users of an expression e are supposed to stand to expert users of e, and in virtue of which the meaning of e out of the mouth of former is supposed to be fixed by its meaning out of the mouth of the latter — can, upon inspection, begin to seem a bit mysterious. The problems here can be presented by way of a kind of dilemma.[20]

---

[18] The examples are from, respectively, Putnam (1973), Burge (1979), and Kripke (1972).

[19] See, among many other places, Horwich (2005), 52-3. [cite others]

[20] Here again I'm following the work of Mark Greenberg, though he shouldn't be held to endorse the arguments which follow. See especially Greenberg (ms.).

On the one hand, deference is often glossed in such a way that the relation is very undemanding — so one might say, for example, that A defers to B with respect to expression e simply in virtue of A intending to use e with the same meaning as B does. But this relation is so undemanding that it does not guarantee asymmetry, and so does not explain why the non-expert's use of e inherits the expert's use, rather than the other way around. Experts, after all, might also have intentions to use the term with the same meaning as non-experts.

To avoid this problem, one might make the relation more demanding, and that A have an appropriately "deferential" attitude toward B — by, for example, being willing to accept correction from B with respect to the use of e. But then it is implausible that we'll be able to cover all of the cases that need to be covered, because A may be, just as a matter of temperament, disposed to accept correction from no one — and an expert might be, again just as a matter of temperament, extremely conflict-averse and concessive when it comes to discussions about the extensions of the terms with respect to which he is expert.[21] And in cases in which no experts exist any more — as in the "Pete Rose" example — it's not obvious what this requirement of being disposed to accept correction from the experts comes to.

So this view of deference is in one way too demanding — but in another way, it is still not demanding enough, as is brought out by the possibility of people being mistaken about who the experts, in the relevant sense, are. Imagine two friends who often go hiking, neither of which knows much about elm trees, but each of which has a very high opinion of the other's botanical knowledge. Each might be inclined to accept correction from the other — indeed, each might intend to use the term "elm" mean whatever his friend means by it — but if we're trying to explain how either gets to use "elm" with its usual meaning, this sort of mutual deference gets us nowhere.

This brings out the fact that there are problems, not just with specifying the content of the deference relation, but also with making sure that the non-experts bear this relation to the right people. And we can't, on pain of circularity, solve this problem by stipulating that the non-experts bear this relation to the experts, and saying that the experts are the people, whoever they are, who use e to mean X — the point of the appeal to deference is to explain the facts in virtue of which the non-experts mean X by e, so we can hardly build in the fact that the relevant experts are the ones who use e with this meaning.

This all makes it tempting to de-emphasize the role of the experts, and say that what's required for deference is just something like the intention to use the term "with its usual meaning in the language." But even this won't solve our problems, since this is an intention which experts and non-experts alike will have, and we'll be left with the problem of trying to say which users play the "expert" role of fixing the meaning of the term in the language. Attempts to do this in terms of who accepts correction from whom will run into the problems discussed above, and attempts to do this in terms of which

[21] Or, as in Burge's example of the word "sofa," the non-expert might resist correction from experts because may be inclined to accept a non-standard theory about the nature of the kind designated by e. See Burge (1986).

speakers satisfy the conditions for "fully grasping" the meaning will be circular. Better, maybe, to just forget about the expert/non-expert distinction, and focus on the question of what it takes to introduce a term with a certain meaning, and what it takes for it to keep that meaning over time.

Another reason why it might be better to think in terms of a theory of meaning maintenance than a theory of deference is that the former leaves open the possibility that a term's maintaining its meaning over time is not a special accomplishment which needs explanation in terms of standing in the deference relation, but rather a default condition which persists unless someone does something about it. This is suggested both by "Pete Rose"-type examples and by our analogy with legislators and laws. In the standard case legislators need do nothing at all to ensure that a law stay in effect; so a theory of "law maintenance" would be a negative theory which would specify what legislators must avoid doing to avoid changing the status of a law. Perhaps the same is true of meaning and speaker actions. But this is so far just a suggestive analogy.

3.4. Perceptual representation and use-properties

The main relevance of the distinction between theories of term introduction and meaning maintenance, for our purposes, is that it seems to open the door for perceptual representation to play the meaning-constituting role in our mentalist theory of meaning. This is because, while it's very implausible to think that using a term with a certain content typically requires any perceptual representation of that content, it's not so implausible to think that *introducing* a term with that content does often depend on a perceptual experience whose content has something in common with that of the term to be introduced.

To see what role perceptual representation might play here, we can begin with a theory which is one of the main alternatives to a mentalist theory of meaning: Paul Horwich's use theory of meaning. According to Horwich's theory, the meaning of a word for an individual speaker *S* is determined, roughly, by that regularity in *S*'s use of the word which best explains *S*'s overall use of the word. We can call these regularities *use-properties*. Some plausible examples of use-properties are:[22]

> "true": the disposition to provisionally accept any instance of the schema, <p> is true iff p.
>
> "and": the disposition to accept any instance of the two-way argument-schema, "p, q // p and q."

The idea is that any speaker whose use of an expression was best explained by, for example, the disposition corresponding to "true" would mean by that expression just what we mean by our term "true." Now, obviously, not all cases will be as easy as the two just listed — but this is enough to get an idea of how one might at least try to proceed.

---

[22] See Horwich (2005), 26-27.

Horwich's theory of meaning is not a mentalist theory, because Horwich thinks that we can always specify use-properties in non-semantic terms. Because by "non-semantic" Horwich means "non-representational" or "non-intentional,"[23] this rules out any appeal to the contents of mental states of users of the language in explaining what gives words their meaning. So why, once we have a theory like Horwich's on the table, why should we need to appeal to facts about perceptual representation?

The reason, I think, is that a theory of Horwich's sort is at its weakest in just the places where perceptual representation can come to its rescue. A good example is the case of proper names. What use use-property constitutes a name's having a certain meaning? Though he emphasizes that the use-property appropriate to a name might vary depending on which type of name we're interested in (names of numbers, for example, might get a different treatment than names of pets) Horwich suggests that the meaning of a demonstratively-introduced name "Pooch" is given by my disposition to accept

> *This* (ostended) puppy is Pooch.[24]

But the worry here is that this sort of use-property pushes the problem back a step, to the problem of giving an account of the meaning of a token of the demonstrative "this" — after all, as Horwich says, "the acceptance of specific sentences containing a word provides it with a definite meaning only relative to to particular construals of the remaining words in those sentences."[25]

Suppose that a particular token of "this" is used to pick out a particular object o — Pooch, as it may be. What use-property could give this token of "this" its meaning? One feels a bit inclined to say that it is a disposition to accept

> *This* (ostended) is *this* (ostending more emphatically)!

but that clearly won't help. A better idea is that the relevant use-property would be something like one of the following:

> The token of "this" means what it does in virtue of my intending to single out o by my utterance of the token.

---

[23] See Horwich (2005), 37.

[24] Horwich (2005), 140; see also the discussion of "Aristotle" in Horwich (1998), 129. An interesting consequence of this sort of view is that it seems to entail that we won't get the differences in the meanings of names which many Fregeans would want, since it it is hard to see how distinct names for an object could be governed by different use-properties if they were both introduced by ostension. But we can, of course, generate standard Frege's puzzle-type cases using such a pair of names.

[25] Horwich (2005), 53. This leaves open the possibility that a pair of words can be interdefined in terms of certain use-properties; this would be an instance of what Horwich calls a "limited holism" or "meaning interdependence." But this would not be a plausible view of the relationship between names and demonstratives, since the latter can be used with or without the introduction of a coreferential name.

> The token of "this" means what it does in virtue of my consciously attending to *o* while uttering the token.
>
> The token of "this" means what it does in virtue of my being disposed to accept "This is *F*" in response to a visual experience which represents *o* as *F*.

One might of course make various objections to any of these. But the problem common to all three is that each violates Horwich's strictures on the inclusion of representational information in the statement of the relevant use-properties. And it is hard (for me at least) to see how to give an account of the use-property for a token of "this" which doesn't have this characteristic.

The nature of these use-properties implies that tokens of demonstratives (and hence also, indirectly, demonstratively introduced names) acquire their meaning from the content of some other intentional state, which means that a mentalist theory for terms such as this is (contra Horwich's non-mentalist use theory) correct. I suggest that a plausible candidate for that other intentional state, from which tokens of demonstratives acquire their content, is a perceptual state which represents the object or property demonstrated. This will then play the role of the meaning-constituting mental state, and we might select various other states — like conscious attention, or intention, or both— to play the role of the linking mental states which ensure that (part of) the content of the meaning-constituting mental state is given to the relevant expression.[26]

Similar remarks might be made about natural kind terms introduced via perceptual demonstrative, as in "Let's call *that stuff* 'water'." Presumably the idea is that the meaning of "water", as in the case of "Pooch", will be given by the use-property which is my being disposed to accept

> *That stuff* (ostended) is water.

But then we're again forced back to the question of which use-property is appropriate to demonstratives like "that stuff." And again it seems as if it will be hard to avoid ending up with a use-property which makes use of representational facts; and, as above, I'd suggest that a plausible candidate for these representational facts will be facts about the perceptual experiences of the relevant subjects.[27] We might then sketch the relevant use-property as follows:

> The token of "that stuff" means what it does in virtue of my consciously attending to the kind, water, while uttering the token.

---

[26] So this view, like Grice's, might analyze meanings partly in terms of the intentions of language users. But here the intentions are playing the role of linking rather than meaning-constituting mental states — and these are not the communicative intentions on which Grice focuses.

[27] This assumes the controversial thesis that natural kind properties are at least sometimes represented in visual experience. For different ways of defending this thesis, see Siegel (2006) and Speaks (2009), which develops an argument of Johnston (2004).

which explicitly involves intentional facts about the objects of states of conscious attention, and implicitly involves facts about the contents of perceptual experience which make the relevant kinds available for conscious attention, and then demonstration.

So one worry about Horwich's treatment of demonstratives and the expressions they are used to introduce is his reluctance to make use of facts about perceptual representation. A separate worry comes from the fact that Horwich does not present his theory as a theory of term introduction, in the sense of the last section, but rather as a full theory of meaning (modulo the need to make exceptions for deferential users). This aim is what leads Horwich to require that use-properties not just be properties of someone's use of the word when introducing it, but also the properties which best explain my use of the name over the full course of the time that I use the name with that meaning. The problem, though, is that it is hard to see how the use-properties associated with demonstrative introductions of names and kind terms could play this role.

A familiar point from Kripke is that my using "Pooch" as a name for a particular dog is compatible with my misapplying the name repeatedly in the future. In particular, it is compatible with Pooch being replaced early in life with a superficially similar dog, Faux-Pooch. How could my acceptance of "*This* is Pooch" at the initial dubbing explain my many later mistaken uses of "Pooch" to describe the exploits Faux-Pooch? And these uses are mistakes — I could be informed of the dog-swap by being told "The dog you've been calling 'Pooch' isn't Pooch."

One might say that the demonstrative use-property explains these later uses when conjoined with the fact that I mistakenly believe that Faux-Pooch is *that dog* (the one I ostended). But the worry here is that this use-property seems less central to explaining my overall use of "Pooch" than does another use-property which governs my use of the word, namely "Pooch is my dog" — after all, the latter explains my mistaken uses of "Pooch" without the extra information about my mistaken belief. But this use-property would entail that "Pooch" out of my mouth has the same meaning as the description "my dog" — which, again for reasons familiar from Kripke, it doesn't. If it did, then "Pooch could have failed to be my dog" would be false, not true (the modal argument), "If Pooch exists, then Pooch is my dog" would be a priori rather than a posteriori (the epistemic argument), and "Pooch" after the dog-switch would refer to Faux-Pooch rather than Pooch (the semantic argument).

The division of labor discussed above suggests a response to this sort of worry: perhaps we should think of use-properties as properties of initial uses of a term, which determine the meaning the term is given upon its introduction into the language; but for them to play this role, there's no reason to require that the meaning-fixing use-property be the one which best explains all of my later uses of the term.[28]

If names and natural kind terms lead to similar sorts of problems for Horwich's theory, quite different difficulties are raised by terms for sensible qualities, like "red." Horwich gives the use-property for "red" as

[28] Though maybe facts about which use-properties do best explain my later uses of the term could play some role in a theory of meaning maintenance.

> "red"'s meaning stems from the fact that its law of use is a propensity to accept "That is red" in response to the sort of visual experience normally provoked by observing a clearly red surface.

But there's a problem in determining exactly which class of experiences is denoted by "the sort of visual experience normally provoked by observing a clearly red surface." A natural first thought is that we should specify this class of experiences in terms of the properties of the surfaces the experiences are experiences of + viewing conditions: for example, we might say that it is the type of visual experiences which normally sighted human beings have when they look at objects with such-and-such reflectance properties under such-and-such light. But the phenomena of spectrum shift (discussed in, among other places, Block (1999)) show that there is no such one type of visual experience. Rayleigh matching tests show that there are systematic differences in the phenomenal characters of the experiences of people of different ages, sexes, and races, all of whom count (and presumably should count) as normally sighted.

An alternative would be to try to specify the relevant class of experiences directly in terms of a certain phenomenal character — say, the phenomenal character of my experiences of red things:

> "red"'s meaning stems from the fact that its law of use is a propensity to accept "That is red" in response to an experience with a RED phenomenal character.

This would be to appeal to a psychological property — but Horwich's aim is not to avoid appeal to psychological properties in general, but only to representational properties. So we should ask whether the property of having a certain phenomenal character is a representational property, or not.

Suppose first that it is not, and that an experience's having a RED phenomenal character is consistent with it representing various color properties as instantiated.[29] Then it seems as though we could have red/green spectrum inverted subjects who nonetheless agree in their visual representation of the colors of things. Given this, it would presumably be possible for them to speak a common color language — to use "red", for example, to mean the same thing. But the use-property just given would make this impossible, since someone spectrum inverted relative to me won't have the propensity it describes, and will instead have the quite different propensity to accept "That is red" in

[29] Here I'm assuming that if there is a necessary connection between phenomenal properties and representational properties, then there the former are identical to some subset of the latter. I think that this is quite plausible, since otherwise it would be mysterious why there would be this necessary connection between the phenomenal and the representational; but one way in which a defender of the present use-property for "red" might resist this argument is by endorsing the supervenience of the phenomenal on the representational without endorsing any such identity claim.

response to experiences with a GREEN phenomenal character — which will suffice for them meaning something different by "red" than I do.[30]

We would do better here, as in the case of names and kind terms, to appeal directly to the representational content of experience. One way of doing this would be to state the use-property which gives "red" its meaning as the disposition to accept certain instances of

> *That color* is red.

where "that color" is a demonstrative whose meaning is given by

> A token of "that color" means what it does in virtue of (i) my perceptual experience representing some object *o* as red, and (ii) my consciously attending to the color *o* is represented to have while uttering the token.

Here again perceptual representation would play the role of the meaning-constituting mental states, and states of conscious attention and/or intention would be the linking mental states.

To this sort of theory, one might object that demonstratives like "that color" typically refer not to the color some object is *represented* as having, but rather to the color the object *does* have.[31] And in this case, why should clause (i) of our use-property talk about perceptual representation at all? Why not just change it to: I'm looking at some object which is in fact red?

One problem here is that in scenarios in which nothing the subject is looking at is in fact red — either because color irrealism is true, or because it is a "brain in the vat" type scenario in which the subject's experiences are all hallucinatory — this sort of use-property would make it impossible to speak about the colors. But this view of the meaning of demonstrative phrases like "that color" is also insufficiently general, for a reason which Mark Johnston has emphasized: hallucinatory experiences can be a source of "original acquaintance" with color properties.

Consider Johnson's example of super-saturated red:

---

[30] Another sort of worry is based on the phenomenon of color constancy, in which we represent the color of a surface as constant through changes in lighting, and hence also through changes in phenomenal character. This, and related cases of color contrast, can make it hard to see exactly what a RED phenomenal character is supposed to be. We could try to get around this problem by just fixing on a single determinate phenomenal character — say, the phenomenal character of my experience of a newly painted fire engine in normal mid-summer sunlight — and specify the use-property for "red" in terms of responses to experiences with *that* phenomenal character. But suppose we introduce a name, "definite-red," for the particular shade of red which such experiences represent a surface as having. It's hard to avoid thinking that the use-property which gives that term it's meaning will be the same as the use-property we've arrived at for "red" — which would make a color term synonymous with a term for a particular shade of that color.

[31] This point is made, in a different context, by Heck (2000), 494.

> "There is a state that a subject can get into by being exposed to bright monochromatic unique green light ... in an otherwise dark room for about twenty minutes. If we then turn the stimulus off, illuminate the room, and have the subject look at a small, not-too-bright achromatic surface, he will see a red afterimage. If the subject turns so that the afterimage is then superimposed on a small red background then something wonderful happens. The subject will then be after-imaging a *supersaturated* red, a red more saturated than any surface red one can see, a red purer than the purest spectral red light ... Supersaturated red is a missing shade of red, which you can only after-image"

We could both be enjoying an experience of this sort, and decide to discuss our experiences, using "that color" as demonstrative term referring to the color our after-image experience represents the surface as having. Here, given our intentions, "that color" will refer to super-saturated red, not to the color that the wall really has — which indicates that there is an intelligible use of demonstrative phrases of this sort on which they are used to pick out the color something is represented as having, rather than the color it has.[32] This is the sort demonstrative for which the use-property given above is appropriate, and it is the sort of demonstrative in terms of which the use-property for "red" and other color words should be given.

The moral of each of these examples — names, kind terms, demonstratives, and terms for "sensible qualities" like the colors — is that a Horwichian use theory needs to make use of facts about perceptual representation. Just so, a mentalist theory of meaning which tries to explain meaning in terms of perceptual representation needs a theory like Horwich's use theory. After all, very many expressions in natural languages — predicates like "is honest" or "is the square root of", connectives like "or", and sentence operators like "necessarily" — have contents which don't figure in the contents of perceptual experience.[33]

To be sure, difficult questions for the friend of this sort of theory remain. One might wonder how we can, from a single use-property, separate out the contributions of character and content to the determination of reference and truth-value.[34] And one might wonder whether Horwich's model can be expanded to give an account not just of the

---

[32] There are also less recondite examples of this phenomenon. Suppose that we're shopping for pants in a store with unusual lighting. We might admire the shade that some pair of pants appears to have in that lighting, and say, "I wish these pants really were *this color*." Here, again, the demonstrative picks out the represented property rather than the property the thing has.

[33] Though I'd qualify this by saying that — for the reasons discussed in connection with the deference relation — even in these cases, use-properties are better suited for use in a theory of term introduction than in a more ambitious theory about what it is for a word to have a meaning for a speaker at a time.

For a defense of the contrary idea that we can give a direct perceptual account of the meanings of these sorts of terms, see Ch. 7 of Prinz (2002).

[34] For a brief discussion, see Swanson (2009).

characters of natural language expressions, but also of the semantic significance of various syntactic constructions.[35]

But I think that a Horwichian use theory, supplemented with facts about the contents of perceptual experiences, has considerable promise. It avoids the problems which arise for mentalist theories which let beliefs or communicative intentions be the meaning-constituting mental states, and, in addition, it validates the (I think, plausible) empiricist thought that our ability to represent the world in language must, ultimately be grounded at least in part in our ability to represent the world in perception.

4. An obstacle to the analysis of perceptual representation

Let's suppose that perceptual states are at least among the meaning-constituting mental states to which stage 1 of our theory of intentionality appeals. How might this affect the prospects of successfully executing stage 2, in which we give a theory of the meaning-constituting mental states themselves?

The usual way of carrying out stage 2 of a mentalist theory of meaning is to give a functionalist theory of the mental states which were the analysans in stage 1. However, some special features of perceptual experience seem to rule out many of the standard ways of constructing such theories.

The following are (sketches of) well-known functionalist theories, as applied to perceptual experience:

Indication theories	the content of a perceptual state is determined by what would cause the subject to be in that state, were the subject in optimal conditions.
Asymmetric dependence theories	a mental representation represents a property if the law L connecting instantiations of the property with tokenings of the representation is such that all other laws governing tokenings of that representation depend on L, and L depends on none of them.
Teleological theories	the content of a perceptual state is determined by the state of affairs which is such that the state was evolutionarily advantageous because of its being triggered by that state of affairs.

It is well-known that externalist theses of the sort listed above are inconsistent with the conjunction of two theses about perception, the first of which is the intentionalist thesis that phenomenal character supervenes on content:[36]

[35] For an interesting discussion of the need for such an account, see King (2007).

[36] Though often, as in Byrne and Tye (2006), the "intentionalism" under discussion includes a the claim that perceptual content supervenes on phenomenal character as well as the claim that character supervenes on content. See Egan and John (ms.) for a demonstration of the incompatibility using the one-direction intentionalist supervenience thesis like that given above.

> *Intentionalism*: Necessarily, any two perceptual experiences with the same content also have the same phenomenal character.

The second is an internalist thesis about phenomenal character:

> *Internalism about phenomenal character*: the phenomenal character of a subject's experiences supervenes on the subject's intrinsic properties.

Both of these theses are very controversial. What I want to show is that we can use considerably weaker, less controversial versions of each, and still get a plausible argument against our externalist functionalist theories — though, as we'll see, it's an argument which depends on some questionable modal assumptions.

Rather than assuming intentionalism, I will make the weaker assumption that the following scenario is impossible:

Psychedelic phenomenology + constant representation of color properties
A subject is looking intently at a well-lit surface which occupies the whole of the subject's visual field. Over the course of a few seconds, his experience goes from being (as we would put it, were we to describe the phenomenal character of the experience) bright-red-feeling (BRIGHT RED, for short) to being BRIGHT GREEN to being BRIGHT RED, and constantly repeats this pattern. And the subject's memory is working normally — it's working pretty much the way yours usually does when you have an experience lasting a few seconds. But, the whole time, he is visually representing the wall as red; it visually seems to him throughout that the wall is red; according to his experience, the wall is red throughout.

This assumption is weaker than intentionalism, because it assumes not that any difference in the phenomenal character of any two color experiences of any two possible subjects suffices for a difference in content, but only that arbitrary variance in the phenomenal character of consecutive color experiences of a single subject does. This thesis is, I think, hard to deny.

Now consider internalism about phenomenal character. Some people find internalist theses of this sort too obvious to need argument; according to John Hawthorne, for example, the truth of this supervenience thesis is a "very obvious fact" and if we "deny this connection between consciousness and intrinsicality ... our very handle on the notion of intrinsicality (and the coordinate notion of duplication) may be thrown into doubt."[37]

[37] Hawthorne (2004), p. XXX.

But others see the thesis as less than obvious — even as "the last bastion of a widely discredited Cartesian conception of the mind."[38]

For the record, I'm in the first camp. However, it is not at all easy to see how to argue for this thesis. It is very tempting to argue from the following intuition:

> "surely we must look to the brain or some such bodily system to make changes in phenomenal character. ... How could the environment effect a phenomenal change in a subject without effecting a brain/body change? To think that it could would be to think that the environment can alter the phenomenal character of experience by causally bypassing the physical mechanisms that make experience possible. It seems extremely plausible that, if you want to change the phenomenal character of someone's experience, you have to mess with their brain. *Just* messing with their environment (in a way that doesn't induce any changes inside their head) won't do the trick."[39]

But this sort of argument seems to show too much. Granted, it is true that I can't make the phenomenal character of the subject's experience change without doing something to his brain. But isn't that also true of a subject sitting around thinking about water? It seems clear that I could not change the topic of that subject's thought without doing something that causally affects his brain — but that doesn't show that the identity of a subject's thoughts are either metaphysically or nomologically necessitated by the subject's intrinsic properties.

So let's set this strong internalist thesis to the side, and instead take a weaker thesis of the counterfactual dependence of phenomenal character on intrinsic properties as our starting point.[40] Let's call this phenomenal dependence:

> *Phenomenal dependence*: the phenomenal character of a subject's experiences counterfactually depends on the subject's intrinsic properties.

---

[38] Byrne and Tye (2006), 242. It may be worth noting, though, that the intuition behind phenomenal internalism is strong enough that it leads many who deny this thesis to say things which would only really make sense if they knew the thesis to be true. Tye's discussion of the example of Swampman in Tye (2000) is a good example. Tye considers the possibility of a molecule-for-molecule Swamp-duplicate of himself wearing spectrum-inverting lenses, and wonders what the phenomenal character of his experience when looking at the sky would be. If phenomenal internalism were false, then it seems as though, given this description of the case, we could reach no determinate verdict about the phenomenal character of Swampman's experience. (This is what an externalist about the content of thought would say about the contents of Swampman's thoughts.) But Tye doesn't say this; he says: "The answer clearly is that ... [it] would have looked yellow" (138). But how could Tye know this, if it is not inferred from an internalist thesis about phenomenal character?

[39] See Egan and John (ms.).

[40] This is "weak C-dependence" in the terminology of Byrne and Tye (2006), who regard the thesis (correctly, I think) as "entirely uncontroversial" (247).

We can then argue as follows. By phenomenal dependence, there is a stable correlation between the internal states of a single subjects and the phenomenal character of that subject's experiences. Let's use S_{RED} as a non-rigidly designating term for "the physical state of a subject which underlies her experiences with phenomenal character RED", and analogously for S_{GREEN}. To argue against that our three externalist functionalist theories entail the possibility of psychedelic phenomenology + constant content we need, respectively, the following claims about possibility:

(a) Possibly, there is a subject for whom S_{RED} and S_{GREEN} both indicate the color red.
(b) Possibly, there is a subject for whom tokenings of S_{RED} and S_{GREEN} both asymmetrically depend upon the color red.
(c) Possibly, there is a subject for whom both S_{RED} and S_{GREEN} were selected for thanks to their being triggered by the color red.[41]

If (a)-(c) are true, we then imagine the subject whose possibility is guaranteed by the relevant claim as alternating between the two states S_{RED} and S_{GREEN}. Given phenomenal dependence, the phenomenal character of his experience would rapidly switch back and forth between RED and GREEN; and given the truth of the relevant externalist functionalist theory, he would be representing the scene before him as red throughout. But this is to say that these theories entail the possibility of psychedelic phenomenology + constant representation of color properties; so they entail something false.[42]

Claims (a)-(c) are certainly not obvious; and to give a really convincing argument for them, we'd need to spell out the corresponding functionalist theories in much more detail. But I think that a plausible case can be made for each of them. Here I'll focus on (a), which I think is the hardest of these modal claims to defend:

> For a pair of states x and y to indicate the same property F (or state of affairs) for a subject is for it to be the case that, in optimal conditions, the subject would be caused to be in x only by instances of F, and the same for y.[43] Optimal conditions are, intuitively, conditions under which the cognitive system of the subject is functioning perfectly. So the possibility of (a) turns on this question: is it possible that a subject could be such that conditions optimal for it are

[41] If we assumed internalism about phenomenal character rather than merely phenomenal dependence, we could weaken (a)-(c) to assert only that, e.g., it is possible that there be a subject for whom the physical states which, for me, underly experiences with phenomenal characters GREEN and RED should both indicate redness.

[42] For discussion of a related argument which uses internalism and intentionalism to argue against functionalist theories of this sort, see Pautz (2006) and Byrne and Tye (2006). See also Pautz (2010).

[43] For ease of exposition, I'm being loose with state types vs. tokens, though not in a way which affects the argument. To be more precise, we should talk about indication relations as holding between state types and properties, and the causal relations in optimal conditions as holding between tokens of those types (or events of subjects coming to be in tokens of those types) and instances of the relevant property.

conditions in which instances of F sometimes cause x and sometimes y, and nothing else ever causes a state of either type?

But this does seem to be possible. Imagine first that our subject is an evolved creature, and that we think of optimal conditions for such creatures as "involving the various components of the visual system operating as they were designed to do in the sort of external environment in which they were designed to operate."[44] It is well-known that the process of evolution often issues in creatures who are not "put together" in the way one might expect a creature designed *ex nihilo* for the relevant environment to be put together. As Gould put the point in a discussion of orchids,

> "Orchids manufacture their intricate components from the common components of ordinary flowers, parts usually fitted for very different functions. If God had designed a beautiful machine to reflect his wisdom and power, surely he would not have used a collection of parts generally fashioned for other purposes. Orchids were not made by an ideal engineer; they are jury-rigged from a limited set of available components."[45]

This sort of observation makes it very plausible that, even if it seems odd a priori, it is possible for a creature to have evolved to operate in an environment in which both S_{RED} and S_{GREEN} were caused by red surfaces. Perhaps there was a surprising absence of green things in the environment, and red things were very important to the creature's survival — so important that it was evolutionarily useful to have redundancy in one's red-indicating capacities.

One might get around this problem by giving another, non-evolutionary view of optimal conditions. But it is hard for views of this sort to steer a middle course between circularity and an implausible indeterminacy of content. On the one hand, we can't define optimal conditions as, e.g., "conditions under which all of the creatures beliefs are true," since that builds facts about the truth conditions of beliefs into the story about what fixes the contents of the subject's mental representations. One the other hand, one can't fall back on a view of optimal conditions according to which they are something like "conditions under which the creature would flourish", since it seems plausible that for at least some creatures, they would flourish as well in a scenario in which S_{RED} was caused by greenness as (with appropriate changes to the causes of other phenomenal states) in a scenario in which it is caused by redness. But it is implausible to think that,

[44] Tye (2000), 138.

[45] Gould (1980), 20.

> for this reason, that creature's perceptual states with phenomenal character RED represent surfaces as indeterminately red-or-green.

This form of argument suggests a constraint on functionalist theories: we should design those theories so that the relation R between mental representations and properties which is such that a representation bears R to a property iff it represents that property should be such that it is not possible that S_{RED} and S_{GREEN} both stand in R to be the same property. Intuitively, and less precisely: the functionalist theory should not permit the equivalent of co-reference between the mental representations which represent the colors.

As noted above, it is plausible that the functionalist theories sketched above fail to meet this constraint. But it might seem as though it is easy to construct a functionalist theory which does:

Covariational theories	the content of a perceptual state is determined by the state of affairs with which the state would covary, were the subject in optimal conditions.

This blocks the argument of the last section, because the requirement of *co*-variation in optimal conditions, by definition, makes it impossible for a pair of states to represent the same color property.[46]

But consider again the possible situation (a) in which both S_{RED} and S_{GREEN} indicate a single color — say, redness. This would then be a case, if the covariational theory were true, in which neither S_{RED} nor S_{GREEN} would covary with a color in optimal conditions, and hence would be a case in which neither represents a color. But now imagine, what is surely possible, that some other states, which we can call S_{ORANGE} and S_{YELLOW}, do covary with colors in optimal conditions — with orange and yellow, respectively. But in this kind of case, if the covariational theory is to be believed, if the subject is looking at a screen with colors being projected upon it, and her experience switches from yellow phenomenology to orange phenomenology to red phenomenology, what has happened is that the screen first visually seemed yellow to the subject, then visually seemed orange to the subject, and then ceased to seem to have any color at all. This is hard to believe. Surely the switch from orange phenomenology to red phenomenology can't be a switch from representing the relevant surface as having a color to simply failing to represent it as colored at all.

And note that this is not *just* a problem for covariational theories. It is a problem for anyone who begins with a theory, like those discussed above, which entails the possibility of psychedelic phenomenology + constant representation of color properties, and then tries to fix this problem by brute force, by adding an extra condition which (in effect) stipulates that no other state represents the relevant property. So if, for example, an

---

[46] This is the theory defended in Tye (1995) and Tye (2000), but for the extra asymmetric dependence condition added in the latter to rule out perceptual representation of intuitively imperceptible qualities with which the perceptible qualities covary. I don't think that this extra clause matters for present purposes.

asymmetric dependence theory entails the possibility of psychedelic phenomenology + constant content, a theory which says that x represents F iff (i) tokenings of x asymmetrically depend on F and (ii) tokenings of no other mental representation do will run into the problem faced by covariational theories.

What makes the functionalist theories just discussed open to these sorts of objections? Two plausibly suggestions are: their commitment to *externalism* and their commitment to *atomism*. Let's consider how the functionalist might avoid the above objections by giving up one or the other of these commitments.

Consider first a purely internalist functionalist theory. If a functionalist theory has the consequence that any intrinsic duplicates are alike with respect to the representational properties of their internal states, then such a theory will not, obviously, permit the permutation of the representational properties of visual states via the permutation of the relationship between those states and instantiations of the various colors in the world, and hence won't be open to an argument like the ones just given.

The problem is that it is not easy to see how to construct such a theory, since even more 'internalist' versions of functionalism, like conceptual role semantics, typically appeal to versions of one of the four theories sketched above in giving their account of the content of perceptual experience. A good example here is Brian Loar, in *Mind and Meaning*, who says that "Perceptual input conditions are needed to secure uniqueness of systematic role because the internal constraints on beliefs do not individuate them" (65). Loar — despite the "internalist" flavor of much of his theory — goes on to give a kind of indication theory for observational beliefs, which can be criticized on much the same grounds given above for criticizing indication theories of perceptual content.

Let's set this sort of theory to the side, and ask whether we might avoid the above line of objection by constructing a theory which abandons the atomist assumption that the contents of certain perceptual state types are fixed independently of others of the same sort. Such a theory might be constructed by analogy with the Fregean theory of perceptual content defended by David Chalmers in "Perception and the Fall from Eden." Chalmers, in discussing the relationship between the phenomenal characters of experiences and the color properties they represent, suggests that this be understood as a "holistic relation ... one can say that the set of [represented color properties] is that three-dimensional manifold of properties that serves as the normal causal basis for the associated three-dimensional manifold of phenomenal .... properties."[47]

The basic idea here is that the color property represented by a given perceptual state is not just determined by the normal causes of experiences of that type — it is given by the one-to-one pairing of color properties and phenomenal types which does the best job of matching phenomenal types with the color properties which are their normal causes. Though Chalmers is not here defending functionalism, there's no reason why a functionalist can't make use of his idea. This would be a functionalist theory which is externalist — since the properties which are the normal causes of experiences might differ between intrinsic duplicates — but would guarantee that the relationship between phenomenal character and represented color properties would be one-to-one, thus ruling

[47] Chalmers (2006), 95.

out the possibility of psychedelic phenomenology + constant content. And it gets this result via holism rather than via a strengthening of the required causal connection — which avoids the problems, discussed above, with the covariational theory.

However, this theory may not be able to avoid entailing the possibility of psychedelic phenomenology + constant representation of color properties after all. However, exactly, we determine which pairing of color properties with phenomenal states does the best job of matching state types with their normal causes, it appears that this pairing is something that can change over time. At least the most obvious definitions of "normal cause" — in terms of frequency, or causation in certain ideal conditions — will have the consequence that the normal causes of a state can change over time, which will be sufficient to change the best pairing of states and represented colors.

Now let t be the time in some individual's life when the "best pairing" changes. Then there will be some pair of distinct phenomenal characters, *c1* and *c2*, which are such that the color property which experiences with *c1* represent before t = the color property which experiences with *c2* represent after *t*. Now suppose that the individual in question is having an experience with *c1* just before *t*, and then, at *t*, switches to having an experience with *c2*. This will then be a case — less dramatic than those described above, but fundamentally no different — of psychedelic phenomenology + constant representation of color properties.[48] Varying the case so that the subject has *c1* before and immediately after *t* would give us a case of constant phenomenology + psychedelic representation of color properties.

Objection 1: Perhaps we should say that there is no "one moment" when the best pairing switches; perhaps the best pairing can only shift over a sufficiently long interval of time *t*.

> *Reply*: Consider two experiences of a single subject, e_1 and e_2, which are separated by the minimal time interval *t*. Because they are separated by *t*, it is possible that they differ in color phenomenology, but have the same color content; to fix ideas let us suppose that at the time of e_1 the subject is such that RED experiences represent the property red, and GREEN experiences represent the property green, whereas in e_2 the subject is such that RED experiences represent the property green, and GREEN experiences represent the property red.
>
> But presumably it is possible for the subject to have a perceptual experience, e^*, during *t*, which must have some color phenomenology — let us suppose that e^* has the phenomenal character RED. What is the content of e^*? Since, by hypothesis, *t* is the minimal interval of time by which two experiences alike in color content but distinct in color phenomenology must be separated, e^* cannot represent the color red, since it is separated from e_2 by an interval less than *t*; and because it is also separated from e_1 by an interval less than *t*, it cannot represent the property green. And e^* can't have some third sort of content since, by varying the description of e_1 or e_2, we could again generate a violation of the

[48] One could stipulate that the "switch" can only happen during an experienceless interval; but see the discussion of the problems with the analogous move to save the time constraint above.

stipulation that t is the minimal interval of time by which two experiences alike in color content but distinct in color phenomenology must be separated.

One might say that it is indeterminate which color property is represented by e*; but it's not clear that this makes much sense. How could an experience with phenomenal character RED represent an object as of indeterminate color? Would an experience of a partly red and partly green object represent the object as indeterminate all over?

One might reply that the minimal interval over the course of which the best pairing switches must be an *experienceless* interval. But this also runs into some difficulties. How long must the experienceless interval be? If it is is relatively short — say, a few seconds — then it seems implausible that such an interval could really make a difference in the contents of the experiences of the relevant subjects for the rest of their lives. And if it is relatively long — say, two hours — then it gets very hard to see how two subjects can differ only in that one has an experienceless interval of two hours and one just a second shorter and yet, from that point on, differ dramatically with respect to whether their visual representation of things as instantiating the property of redness is done via experiences with the phenomenal character RED or the phenomenal character GREEN.

One might be tempted to reply by pointing out that this is just a sorites argument, and so should be open to whatever resolution sorites arguments in general should get. But I think that that would be a mistake. We can't say that "borderline cases" — say, cases where the subject's experienceless interval is between 1:55 and 2:05 — are ones in which the content of the subject's experience after the interval is "indeterminate," since (as above) it is very implausible that it can be indeterminate whether a visual experience of any phenomenal character represents an object as red or green. (Again, imagine a visual experience of something which is half-red and half-green — would the experience represent the object as indeterminately-red-or-green all over?) So we're forced to say that there must be a sharp cut-off point here — and, whatever the attractions of epistemic views of vagueness in other cases, it seems very unattractive here. For one thing, it looks like the sharp cut-off point here would be (as in other cases of vagueness, on the epistemic view) undiscoverable, which in turn would make it hard to see how certain subjects could be in a position to know what which properties their own current visual experiences would be representing objects as having.

Objection 2: maybe the "normal cause" of a state does vary over time; but nothing says that we have to define the pairing between the phenomenal states and the colors in terms of normal cause. And the argument just given seems to show that we shouldn't, and that we should instead let the pairing be determined by some fact about the subject which, in principle, can't change over the course of the subject's life — like, for example, the subject's evolutionary history.

Reply 1: One consequence of this sort of view is that it seems to make interpersonal spectrum inversion (and spectrum shift) without misrepresentation possible, but intrapersonal spectrum inversion (and spectrum shift) without misrepresentation impossible. This can seem ad hoc, since the intuitions which militate in favor of the possibility of interpersonal inversion are just as strong in the intrapersonal case. A good example here is the sort of spectrum shift discussed in Block (1999). We're just as disinclined, intuitively, to assert differences in the veridicality of color experiences between the experiences of people of different ages (intrapersonal shift) as between the experiences of people of different sexes and races (interpersonal shifts).

But this might be taken just to be an interesting consequence of, rather than an objection to, the view.

Reply 2: A holist teleological theory of this sort can't be the last word, since it is extremely plausible that Swampman could visually represent objects in his environment as having colors. So to make this response general, we also need some account of what could fix the correlation between Swampman's phenomenal states and the colors which is such that the "best pairing" can't change over the course of Swampman's life. One might appeal to some more general understanding of "optimal conditions" to do this, but this gets into the problems discussed above in connection with modal claim (a) — it is hard to give such a view of optimal conditions without either lapsing into circularity or giving rise to an implausible indeterminacy of content.

If this sort of worry can be overcome, then it seems to me that giving some theory of this sort — a holist functionalist theory which fixes the pairing between phenomenal states and represented properties using some property which is in principle unchangeable over the course of the subject's life — may be the best bet for the externalist functionalist about perceptual content.

5. Conclusion

The idea that we can base a mentalist theory of meaning on perception rather than belief or intention has some interesting features. Of particular interest, I think, is that it captures some of the central intuitions behind both "mind-first" and "language-first" approaches to intentionality.

The first is pretty obvious: the view just sketched shares with "mind-first" approaches the view that a certain species of mental representation — in this case, perceptual representation — is a kind of "original intentionality" in terms of which linguistic representation must, ultimately, be explained.

But it also shares some important things with "language-first" views. Consider the following expression of this sort of view of intentionality by Michael Dummett:

> "The fundamental point that language enables us to grasp new thoughts is often sneered at by those too impatient to reflect upon it. ... If it was merely that we could understand a new way of conveying a familiar thought, a language could be simply a code for thoughts. ... This picture of language depends on taking our grasp of the thoughts that can be conveyed by language as antecedent to our understanding of the language: it is therefore exposed as false by the fact that such an understanding suffices to enable us to grasp quite new thoughts when they are expressed in that language."[49]

Dummett's emphasis here is not on the global relationship between mental and linguistic representation: it is on the order of explanation, for a particular thinker, between understanding a sentence which expresses a certain proposition and that thinker's ability to have thoughts with that proposition as their content.

On this topic — the topic of the relationship between thought and belief, on the one hand, and linguistic meaning, on the other — the present view comes out very much on the side of Dummett. After all, on this sort of view, it might often be the case that language is genuinely a vehicle for the thoughts of a subject, in the sense that it will be — given the subject's location in space and time — impossible for the subject to think the relevant thought without the mediation of a sentence of her language expressing that thought. And it even leaves open the possibility of the stronger claim that an analysis of what it is for a subject to believe or judge a certain proposition will have to make use of facts about the meanings of expressions in a public language.[50]

There's thus a sense in which this view both takes mental representation to be, ultimately, more fundamental than linguistic representation, and in many cases gives language explanatory priority over thought and belief.[51]

Bibliography

Block, Ned. 1999. "Sexism, Racism, Ageism and the Nature of Consciousness," *Philosophical Topics*, 26.

Burge, Tyler. 1979. "Individualism and the Mental," *Midwest Studies in Philosophy*, 4: 73-121.

———. 1986. "Intellectual Norms and the Foundations of Mind," *Journal of Philosophy*, 83:12: 697-720.

———. 1975. "On Knowledge and Convention," *Philosophical Review*, 84:2: 249-55.

Byrne, Alex, and Michael Tye. 2006. "Qualia Ain't in the Head," *No\ ^us*, 40: 241-55.

Caplan, Ben. 2005. "Against Widescopism," *Philosophical Studies*, 125: 167-90.

---

[49] Dummett (1989), 173.

[50] I defend this stronger claim in Speaks (forthcoming).

[51] Thanks to participants in my graduate seminar at Notre Dame in the fall of 2009, and an audience at the Colloquium on Thought and Language at the University of Cincinnati for helpful discussion of this material.

Chalmers, David. 2006. "Perception and the Fall from Eden". *Perceptual Experience*. New York: Oxford University Press.

Dummett, Michael. 1989. "Language and Communication". *The Seas of Language*. Oxford: Oxford University Press.

Egan, Andy, and James John. ms. "A Puzzle About Perception."

Evans, Gareth. 1973. "The Causal Theory of Names," *Proceedings of the Aristotelian Society*: 187-208.

Gould, Stephen Jay. 1980. *The Panda's Thumb*. New York: W. W. Norton & Company.

Greenberg, Mark. ms. "Incomplete Understanding, Deference, and the Content of Thought."

Grice, Paul. 1957. "Meaning," *Philosophical Review*, 66:3: 177-88.

———. 1969. "Utterer's Meaning and Intentions," *Philosophical Review*, 78:2: 147-77.

———. 1968. "Utterer's Meaning, Sentence Meaning, and Word Meaning," *Foundations of Language*, 4: 225-42.

Hawthorne, John. 1990. "A Note on `Languages and Language'," *Australasian Journal of Philosophy*, 68:1: 116-18.

———. 2004. "Why Humeans Are out of Their Minds," *Nous*, 38: 351-58.

Heck, Richard. 2000. "Nonconceptual Content and the ``Space of Reasons'',"
Philosophical Review, 109:4: 483-524.

Horwich, Paul. 1998. *Meaning*. Oxford: Clarendon Press.

———. 2005. *Reflections on Meaning*. Oxford: Clarendon Press.

Johnston, Mark. 2004. "The Obscure Object of Hallucination," *Philosophical Studies*, 120: 113-83.

King, Jeffrey C. 2007. *The Nature and Structure of Content*. New York: Oxford University Press.

Kripke, Saul. 1972. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, David. 1975. "Languages and Language".

———. 1992. "Meaning without Use: Reply to Hawthorne," *Australasian Journal of Philosophy*, 70:1: 106-10.

Loar, Brian. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.

Neale, Stephen. 1992. "Paul Grice and the Philosophy of Language," *Linguistics & Philosophy*, 15:5: 509-59.

O'Leary-Hawthorne, John. 1993. "Meaning and Evidence: A Reply to Lewis," *Australasian Journal of Philosophy*, 71:2: 206-11.

Pautz, Adam. 2010. "Do Theories of Consciousness Rest on a Mistake?," *Philosophical Issues*, 20: 333-67.

———. 2006. "Sensory Awareness Is Not a Wide Physical Relation: An Empirical Argument against Externalist Intentionalism," *No\^us*, 40: 205-40.

Prinz, Jesse J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Putnam, Hilary. 1973. "Meaning and Reference," *Journal of Philosophy*, 70: 699-711.

Schiffer, Stephen. 1972. *Meaning*. Oxford: Oxford University Press.

Siegel, Susanna. 2006. "Which Properties Are Represented in Perception?". *Perceptual Experience*. New York: Oxford University Press.

Soames, Scott. 2002. *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford: Oxford University Press.

Speaks, Jeff. forthcoming. "Explaining the Disquotational Principle," *Canadian Journal of Philosophy*.

———. 2009. "Transparency, Intentionalism, and the Nature of Perceptual Content," *Philosophy and Phenomenological Research*, 79: 539-73.

Swanson, Eric. 2009. "Review of *Reflections on Meaning*," *Philosophical Review*, 118: 131-34.

Traugott, Elizabeth C., and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Tye, Michael. 2000. *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

———. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations. Translated by G. E. M. Anscombe*. New York: MacMillan.

———. 1937-1944. *Remarks on the Foundations of Mathematics. Edited by G.E.M. Anscombe, G.H. Von Wright, and Rush Rhees. Translated by G.E.M. Anscombe.* New York: MacMillan, 1994.