Chapter 5

Deflationism about truth and meaning

(Originally published in *Southern Journal of Philosophy* 40: 217–242, 2002)

Preview

Some philosophers (e.g., Paul Horwich 1998) argue that we should be deflationists about meaning. Such deflationism is an interesting new position in the debate about meaning. However, though deflationism about meaning has some attractive features, it will not be successful in giving an adequate account of meaning. The theory is intended to be deflationary in the sense that it is parallel to, and justified by, deflationism about truth. I show, focusing on Horwich's very detailed notion of semantic deflationism, that it is not particularly parallel to deflationism about truth, and neither is it in any adequate sense justified by deflationism about truth.

Chapter 5

Deflationism about truth and meaning

0. Introduction

Deflationism is an extremely influential approach to the philosophical problems surrounding the concept of truth. It is the theory, basically, that there is nothing more to truth than can be explained by instances of the schema 'the proposition that *P* is true if and only if *P*'. Recently some philosophers have proposed to extend the deflationary strategy to the debate about the concept of meaning. Thus Paul Horwich has proposed a deflationary use theory of meaning as an integral part of an overall position called *semantic deflationism* which also comprises deflationism about the truth-theoretical terms 'true', 'true of' and 'refers'.[1]

Others, for example Schiffer 1987, Johnston 1988, and Field 1994, have, in a more meaning sceptical frame of mind, advocated various deflationary and minimalist approaches to meaning. Horwich's approach differs from these, not only because his theory is by far the most detailed and substantial defense of this new position in the philosophy of language, but also because he is not a meaning sceptic. He argues that, viewed in the right deflationary light, a use theory can give us a reductive account of meaning, that is, an account of meaning in non-normative, non-semantic terms.

The concept of meaning has proven very resistant to philosophical treatment. No-one has come up with plausible naturalistic analyses of meaning, and some people even argue that no facts of any kind—naturalistic or otherwise—could make true our statements about

meaning.[2] Obviously, it would be very helpful if we could employ deflationism to make inroads on this seemingly intractable problem of meaning. Indeed, semantic deflationism would be a very attractive package of solutions to problems about truth, reference and meaning. However, I do not think the strategy is going to be that successful, even if we assume deflationism about truth. Horwich's deflationary use theory of meaning is intended to be deflationary in the sense that it is both parallel to, and justified by, deflationism about truth. After outlining the use theory, I shall argue, in Section II, that the use theory is not in any strong sense parallel to deflationism about truth. And in Section III and IV I show that the attempt to justify the use theory by deflationism about truth does not give us an adequate account of meaning. In particular, in attempting to cash out Horwich's explanatory strategy for three different interpretations of the use theory, we find that either (i) the use theory will not *explain* meaning; or (ii) the use theory will not be *deflationary*; or (iii) the use theory will not be properly *reductive*.

I. Horwich's use-theory of meaning.

The question Horwich is concerned with is: "What is meaning?" That is, how come some inert sounds are imbued with meaning and others are not; how come they can reach out and *mean* some definite portion of the world (p. 1)? Horwich proposes that meaning-properties, e.g., 'x means DOG',[3] are constituted by non-semantic explanatorily fundamental use-properties. The account is construed such that it is parallel to, and justified by, deflationism about truth (p. 5, 41, Ch. 10).

It is important to notice that Horwich does not promise us a reductive analysis of what *meanings* are. The slogan "meaning is use" does not amount to the claim that meanings somehow *are* uses. Rather, the claim is that use constitutes the relation between word and meaning. Thus, for example, it is in virtue of the use of 'wombat' that that word comes to be

related to the meaning WOMBAT in English. Meanings themselves are the concepts typically expressed by tokens of the word. For example, when I say "there is a wombat" that typically indicates the presence in me of a thought with the propositional character THERE IS A WOMBAT. The meaning of the word 'wombat' is then the concept WOMBAT which is an abstract component of such mental states (p. 26-7, 44, 98f).

i) a central problem in the traditional approach: relationality.

We saw that there is a straightforward sense in which meaning is relational. A word's meaning property is that which tells us that use of the word *indicates* the presence, within the mind of the speaker, of a certain concept. But this aspect of relationality should not make us assume that the underlying, meaning-*constituting* properties must also be relational. The reason is, as Horwich argues, the general one that relatively superficial relational properties can themselves be constituted by underlying non-relational properties. Constitution of *F*-properties by *G*-properties only requires that *F* and *G* are co-extensional and that facts about *F* are explained by facts about *G*; it does not require that *F* and *G* have the same logical form (p. 24-26). Therefore, from the fact that meaning-properties are relational it does not follow that meaning-constituting properties are relational (though they might be). This has the consequence that, even though we are able to read off from a word's meaning-property that the word indicates a particular concept, we may not be able to glean that information from the underlying, possibly non-relational, meaning-constituting property that makes it true that the word indicates that particular concept.

This view of constitution goes against what many people have thought is essential to an analysis of meaning, viz. that one be able to *read off* which meaning is constituted from the proffered meaning-constituting properties. The notion of 'reading off' is metaphorical, but perhaps we can explain it in terms of supervenience.[4] If the meaning-properties *M* supervene

on the meaning-constituting properties *C*, then the truths about *C*-properties (plus, perhaps, various conceptual truths) entail the truths about *M*-properties. The 'reading off' requirement then comes to the notion that this entailment is *a priori*: that we are able to know the *M*-truths merely by knowing the *C*-truths. Notice that, if one believes that the entailment is *a priori*, then it seems reasonable to insist that the *C*-properties be relational in the sense that they concern relations between word tokens (e.g., tokens of 'dog') and (instances of) properties (e.g., doggyness). How else could we ever read off what the word's meaning is from those underlying properties alone?

This has significance for the wider debate about meaning. For it seems it is the reading off requirement, and therewith the idea that meaning-constituting properties must be relational, that has proven to be the stumbling block for many attempts to naturalize semantics. The problem is that it seems close to impossible to specify a substantial non-semantic relation between a word and an extension such that, without begging the question, the meaning of the word can be *read off*, in the sense explained, from this relation. How for instance do we exclude dog-like sheep from being a member of the extension of the word 'dog' without presupposing that 'dog' is true of *only* dogs? And, what is it about the finite number of occasions of my use of 'dog' that relates it to *all* the possible dogs? This is one of the core problems in the debate about meaning. It is, for example, closely related to Kripke-Wittgenstein meaning-scepticism (Kripke 1982, Ch. 2) about dispositional accounts of rule-following (cf. p. 24 and Ch. 10). Horwich's point is that we can avoid the problem if we reject the requirement that we must be able to read off meaning-properties from the meaning-constituting properties, and that the meaning-constituting properties therefore had better be relational.

Naturally, this leaves Horwich with an explanation problem. Constitution of *F*-properties by *G*-properties requires that facts about *F*-properties be explained by *G*-properties. *A priori* entailment being rejected, we are then owed an explanation of how Horwich's meaning-constituting use-properties can explain the relationality of meaning-properties. We shall see (in Sections III and IV) that, surprisingly, it is an explanation that relies just as much on a species of the notion of knowing-how as on the notion of *a posteriori* knowledge

ii) deflationism and meaning.

Deflationism about truth is the theory which says that all of what there is to say about deployment of the concept of truth can be captured by instances of the equivalence schema: 'the statement [or proposition, or belief] that *P* is true iff *P*'. The reason is that it appears the concept of truth functions as nothing but a device of generalization. For example, it allows us to generalize on the basis of

(A) If Florence is smiling, then Florence is smiling

to get,

(G) Every statement of the form 'if *p*, then *p*' is true,

even though (A) does not have the form that normally allows generalization. The equivalence schema gives us

(A') The statement *that if Florence is smiling, then Florence is smiling* is true

and thus it gives us a term referring to an object (a statement) to which we can attribute a property, and which we can replace with a universally quantified variable in order to arrive at the generalization. Without the intervention of the schema, (A) is not apt for such generalization.[5] The deflationist's claim is that this type of move is representative of the only function of the concept of truth, and that all it requires is the equivalence schema. Hence, no so-called 'inflationist', substantial property of truth (i.e. correspondence, coherence, verification etc.) is needed.[6]

In what follows I assume deflationism about truth. It is after all a relatively popular theory, and it does explain our usage of 'true' rather well (as witnessed by the battery of arguments in Horwich 1990).[7] Moreover, it would lend further support to deflationism about truth if it could help us to an adequate account of meaning. (The importance of deflationism is emphasized by, as we shall see in Sec. III, Horwich's argument that inflationists about truth *must* rely on the problematic notion of substantially relational meaning-constituting properties).

There is a clear sense in which renouncing the requirement that all meaning-constituting properties must be relational is *parallel* to one of the tenets of deflationism about truth. Everyone agrees that, for example, the sentence 'snow is white' is true iff snow is white; likewise, the sentence 'grass is green' is true iff grass is green. The deflationist about truth points out that snow's being white and grass's being green have nothing interesting in common. Different facts ensure the truth of different sentences. We should not expect there to be anything substantial in common between the truths of the two sentences; in particular, we should not expect the truth of the two sentences to be constituted by the same type of substantial relation.

It is similar in the case of meaning. The meaning of 'wombat' is constituted by a property *U* possessed by the word 'wombat'; the meaning of 'democracy' is constituted by a property *U'* possessed by the word 'democracy', and so on. There are substantial properties that constitute meaning, but the properties *U* and *U'* need have nothing interesting in common; in particular we should not expect the meanings of the two words to be constituted by the same type of substantial relation.[8]

The use-based account of meaning is not only *parallel* to deflationism about truth in the way I have just explained, it is also *justified* by it.[9] I will explain and discuss this aspect of the theory in sections III and IV, where the focus is on how Horwich manages to explain so-called 'aboutness' without relationality.

iii) what are the meaning-constituting properties and how do we identify them?

The property of a word that constitutes its having this or that meaning is *the explanatorily basic, relatively simple, acceptance property* that best explains the overall use of that word in various sentences:

> [F]or each word, *w*, there is a regularity of the form
>
> > All uses of *w* stem from its possession of acceptance property *A(x)*,
>
> where *A(x)* gives the circumstances in which certain specified sentences containing *w* are accepted (p. 45).

Here are some of Horwich's own examples of such basic use regularities:

> The acceptance property that best explains a speaker's overall use of the word "and" is (roughly) his tendency to accept "*p* and *q*" iff he accepts both "*p*" and "*q*".
>
> The explanatorily fundamental acceptance property underlying our overall use of the word "red" is (roughly) the disposition to apply "red" to an observed surface when and only when it is clearly red.
>
> The acceptance property governing our total use of the word "true" is the inclination to accept instances of the schema 'the proposition *that p* is true iff *p*' (p. 45, cf. also p. 129).

Straight off this look fairly plausible, and since the talk is of tendencies, dispositions and inclinations of use it also seems apt to fulfill the important promise of being a *reductive* account of meaning-properties (p. 5).

How do we go about *identifying* these basic acceptance properties, given there may be no relational properties from which we can 'read off' (in the sense explained above) the meaning of the word? Here is what Horwich says:

> Think of all the facts regarding a person's linguistic behavior—the sum of everything he will say, and in what circumstances. The thesis is that this constellation of data may be unified and explained in terms of a relatively small and simple body of factors and principles including, for each word, a basic use regularity (p. 45).

Hence, the *explanandum* consist in the entire set of tokens of the word in question. The linguistic theorist can then, using his or her home language, employ ordinary scientific

criteria (unification, simplicity, predictive power etc.) in arriving at the *explanans*: some basic acceptance property which best explains the overall use.

What I have said so far constitutes the core elements of what I take to be Horwich's account of the non-semantic facts that constitute meaning-properties. There are three parts in what follows. In Section II I focus on the extent to which the use theory is *parallel* to deflationism about truth. In Sections III and IV I turn to a discussion of how, given we have identified the basic acceptance properties, they can be viewed as explanatory and basic.

II. An account of meaning parallel to deflationism about truth?

The first worry one might have about this account is that *acceptance* seems to be a semantic notion (because the sentences we accept are the sentences we hold true). How can the analysis of meaning in terms of acceptance properties be *basic* if these properties are themselves semantically constituted?[10] To allay this worry Horwich points out, firstly, that *acceptance* is not constituted by the notion of truth any more than the notions of *doubting* and *supposing* are. Secondly, he argues that, in so far as there is a link to truth for these notions, it can be fully explained by the equivalence schema. Moreover, if we employ all the resources of psychological explanation we can, Horwich suggests, arrive at good functional theories that allow us to simultaneously characterize the notions of 'acceptance', 'desire', 'observation', and 'action' (p. 95-96). I propose we grant Horwich that we can, on purely psychological-functionalist grounds, get a feasible idea of which sentences speakers accept, or hold true. It is in many ways an old complaint, and Horwich gives us good reasons to believe it can be overcome.[11]

Instead, I want to pursue this notion of a psychological-functionalist methodology. It seems to me that, though we may be able to arrive at the type of explanatory regularities Horwich is

concerned with, there is a diversity of other, apparently complementary, explanations available of each token word in the *explananda*. This raises a question concerning what we may call the explanatory cohesiveness of Horwich's account. Discussion of this question will allow us to appreciate the strong functionalist nature of the theory and, furthermore, evaluate the extent to which it is parallel to deflationism about truth.

Consider the behavior represented by the set of all my tokens of 'poisonous'.[12] Different tokens may be explained by a wide range of different underlying states. Thus, for example, I may utter a token of 'poisonous' in a sentence because I believe that something is poisonous and I desire to warn someone; or I may be in the pub telling a lame joke, so I utter a token of 'poisonous' because I believe that some substance is not poisonous, but I desire to pull someone's leg; or I may utter a token of 'poisonous' because I believe that some substance is poisonous, I believe that my interlocutor is a daredevil and I desire him or her to take the poison, and so on. There are also many more wayward examples: I utter a token of 'poisonous' because I believe that someone wrote me a nasty letter and I desire to let someone know. And so on.

These examples reflect a natural sense in which we can say that our beliefs and desires explain our linguistic behavior. If we want to know why this or that particular token was uttered we cite the underlying beliefs and desires. And as we have seen, different beliefs and desires underlie different tokenings of the word. In general, the underlying states that explain the members of the set of my tokens of 'poisonous' may have very little in common. (This situation is aggravated if we add the holistic nature of belief-desire explanations, in that case any of the regular changes to my beliefs and desires may affect the explanation for a given utterance).

The point of this is to show that, if Horwich is right, then we seem to have two types of explanation of the same phenomena. On the one hand there is the explanation of a token of 'poisonous' in terms of its basic acceptance property. On the other hand there is the explanation of the same token in terms of the beliefs and desires of the subject uttering the word on any particular occasion. These explanations are distinct in the sense that the next token of 'poisonous' can and probably will be explained by different beliefs and desires, but by the same acceptance property.

Clearly we need both types of explanation (assuming it is Horwich's acceptance properties and not some other meaning constituting property we need). When you offer me a piece of cake saying "this cake is poisonous" I have an interest in finding out what your beliefs and desires are, even though I may already know the acceptance property possessed by 'poisonous'. And I may have a reasonable idea of your beliefs and desires, but be puzzled about your choice of words, that is, wonder which acceptance property explains your choice of words.

Horwich clearly allows that basic acceptance properties only *partly* explain tokens of words (p. 47). In other words, a full explanation would have to draw on basic acceptance properties as well as the particular beliefs and desires of a speaker. But whereas it is clear that the explanations ought to combine, it is not so clear *how* they combine.

I want to suggest that the best answer to this question is found by recalling how *functionalism* in general is especially well-suited to handle such diversity problems. Perhaps, that is, we should take Horwich's allusions to functionalism seriously (p. 44, 87, 95, 99), and conceive of basic acceptance properties as functional properties that take perceptual or conceptual content as input, produce tokens of words in accepted sentences (or simply,

acceptances) as output, and which are interrelated with other functional acceptance properties. The problematic fact that there is a heterogeneous set of belief-desire explanations can then be dealt with if we allow different sets of beliefs and desires to *realize* the same functional role on different occasions. In this way different tokens of 'poisonous' may be explained by different sets of beliefs and desires, and yet every token may be explained by the same functional role specified by the basic functional acceptance property for 'poisonous'.

Adopting such a functionalist strategy would allow us to tell how the two types of explanations can be complementary: when we seek commonsense belief-desire explanations we often take the meanings of words for granted, and when we seek explanations of the meanings of words we must rely on what we take people's particular beliefs and desires to be on various occasions.

Compare with this case. On a particular occasion we may want to give an explanation of Jo's opening the fridge. We cite the particular beliefs and desires (say, Jo's desire for a chilled beer, her belief that there is beer in the fridge, and the belief that if she opens the fridge she will get the beer). Next time Jo opens the fridge we may explain it by citing another set of beliefs and desires (say a desire for a carrot, etc.). Yet, in general, different fridge-openings may be explained by the same 'basic regularity': having a desire for something one believes is in the fridge. This regularity may indeed explain all fridge-openings, in spite of the underlying diversity (desires for beer, carrots etc.). In this case it seems reasonable to say that the basic regularity is an overarching functional role realized in different cases by different particular sets of beliefs and desires. My claim is that this is similar to the two different ways we may explain tokens of words.[13]

The close affiliation with functionalism puts us in a better position to evaluate the manner in which the deflationary use theory of meaning is *parallel* to deflationism about truth. When viewed as a functionalist theory we can, after all, allow a familiar *general* conception of the use properties that constitute what words come to mean. Meaning-constituting properties are functional properties identified on the basis of a psychological method employing scientific criteria of unification and simplicity. These functional roles can be realized on different occasions by different sets of beliefs and desires.

This is still in accordance with an account of meaning that in some sense is parallel to deflationism about truth in as much as the functional roles for different word types (of the same syntactic category) need not have much in common (just as smoke alarms and coke machines are defined by different functional roles). However, it exerts some pressure on the extent to which the theory is truly parallel to deflationism. We can grant Horwich that not *all* acceptance properties will be similar. But surely we could still expect different discourses to be internally homogenous, yet quite different from other discourses? (Just as we should expect there to be interesting structural resemblances between the functional roles for smoke alarms and gas detectors, and other interesting structural resemblances between the roles for coke machines and coffee dispensers). Indeed, for some word, e.g., 'red' to belong to a certain discourse, e.g., color discourse, we may find that it is a necessary condition that its basic acceptance property structurally resembles the basic acceptance properties for the other color terms, e.g., 'blue' and 'green'. The potential for such 'regional' structural commonalities goes against the grain of deflationism because it seems unduly pessimistic to hold that we will be unable to read off anything about meaning-properties directly from the meaning-constituting properties in question.

This need not be a too serious problem for Horwich's overall project, but it does question the extent to which the account is really 'parallel' to deflationism about truth; and with that it questions the scope of the attack on the *reading off* requirement and the notion of extensional relationality.

Notice that I am not advocating simply that basic acceptance properties should be conceived as functional properties. I am arguing that, *given* we require an account which make explanations in terms of acceptance properties and belief-desire explanations cohere, we had better conceive of acceptance properties as functional properties. I do suspect the semantic deflationist would prefer acceptance properties to offer *deeper* explanations than the ones offered by overarching functional acceptance properties. Clearly, if semantic deflationism is going to be basic, then the deflationist should be wary of endorsing functional acceptance properties which are realized by contentful beliefs and desires. But, if the deflationist insists that acceptance properties are not such functional properties, then we are at the very least owed an account of how the use theory can be properly cohesive.

In what follows the above themes will re-occur. At the beginning of Section IV I return to the notion of structural similarities for the acceptance properties for terms belonging to color discourse. And towards the end of Section IV I shall provide independent support for the worrying idea that semantically *basic* regularities are themselves realized by seemingly contentful mental states.

III. A Deflationary Explanation of Aboutness.

i) an account of meaning justified *by deflationism about truth.*

Horwich's use-theory is not only intended to be parallel to, but also to be *justified* by, the deflationary theory of truth (Cf. p. 10-11, 42, 69-71, 108-110, 113). I shall outline how I think this story goes, then I will question how this story fits in with the project of giving an explanatorily basic account of aboutness. I shall suggest that there is no readily available notion of acceptance properties which will be explanatory *and* deflationary *and* basic.

ii) meaning and deflationism about truth.

The idea, in short, is as follows. Once the meaning of a word has been specified it should be possible to specify what it is *true of*. For example, once the meaning of 'wombat' is specified, it follows trivially that 'wombat' is true of all and only wombats. If we are inflationists about the notion 'true of', then we must insist on analyzing that notion in relational terms. This is because inflationists hold that the fundamental principle of *being true of* is something like:

Inflated: $(x)\ (y)\ (x$ is true of y iff $Rxy)$

Where '*R*' is some substantial relation. According to Horwich, given the close ties between meaning and the notion of being true of, the meaning-constituting property must then also be substantially relational: otherwise it is hard to see how it could follow trivially from the specification of a word's meaning what it is true of. And this, as I discussed in Sec. I.i., is what creates the meaning sceptical problems associated with the 'reading off' requirement.

The meaning sceptical threat can be avoided if we accept deflationism about the notion ‘true of’. That simply amounts to accepting schemes like this:

Deflated: (*y*) (*F* is true of *y* iff *Fy*)

In this case there simply is no relationality constraint on the ‘true of’-notion, and the meaning-constituting property therefore also remains untainted (Cf. p. 69-71, 108-10).

Since we, according to Horwich’s overall position—semantic deflationism—have good independent reason to accept deflationism about the truth-theoretical notions, we have good reason to believe that meaning-constituting properties do not have to be substantially relational. This is the sense in which the use-based account of meaning is *justified* by deflationism about truth.

iii) explaining aboutness.

But, one might ask, isn’t the relationality constraint imposed for a good reason? Isn’t that the only way we can hope to explain how words are *about* things? How can we get aboutness without being able to *read off* what the word is about from its meaning-constituting property? And how can we begin to read off in this way if the property does not relate the word with the members of its extension? In other words, it might be that extensional relationality is a good thing, and that even the deflationist about truth should insist on it.

Horwich argues that we can account for aboutness without honoring the relationality requirement. If we adopt deflationism, and abolish the relationality requirement, then we can motivate another explanatory strategy:

> [T]here will be no way to read off which meaning is constituted by a given use property. The best we can do, in order to get from one to the other, is to appreciate that some word (say, 'glub') has the use property—i.e. to actually use it in that way; in which case we can deploy that very word to characterize the constituted meaning (as "x means GLUB"). (p. 66; Cf. also p. 89; 107-112).

In other words, semantic deflationists can avoid relationality and yet explain aboutness because they in effect *order* their explanation in two steps:

> *Step 1*. The (potentially non-relational) basic acceptance property *U* constitutes the meaning-property of the word type α.
>
> *Step 2*. If α has the property *U*, *then* α's meaning can be specified by appreciating that it has *U*, that is, by putting oneself in a position to use the word to characterize its meaning.

As I see it, one of Horwich's core claims is that segmenting the explanation of meaning like this frees it of the requirement of substantial relationality and thereby obviates the Kripke-Wittgenstein problems associated with our not being able to read off meaning from meaning-constituting properties. The idea is that this move is only available to deflationists about truth.

This is an interesting and attractive idea. I take it the basic idea is that, from an external perspective (employing the controversial extensional relationality requirement) on linguistic practices, it is easy to see how indeterminacy and multiple interpretations are possible, but from an ordinary insider perspective questions of indeterminacy and interpretability hardly

arise. As such linguistic insiders we are, as it were, entitled to disquote ourselves in a carefree manner. Horwich rightly sees that insisting on extensional relationality is insisting on an external perspective, and therefore tries to block that route by drawing heavily on the services of deflationism about truth.

iv) How should we conceive of acceptance properties in order to explain aboutness?

I want now to raise a question about this two-step explanatory strategy: how should we conceive of the basic acceptance properties, in order to successfully discharge the crucial second step of the explanatory strategy?

It is important to acknowledge that in order for Horwich's semantic deflationism to succeed, the explanatory strategy must rely on step 2 just as it must rely on step 1. It may well be that we can identify some acceptance properties for each of the words of a language (using the functionalist-psychological method outlined in Section II). But obviously we will not have succeeded in explaining aboutness unless we can in addition show what it is for someone to appreciate that a given word has a certain acceptance property, i.e. show what is involved in someone using it in that way to characterize its meaning. If this cannot be adequately done, then we will not, after all, have an explanation of aboutness. That is, if we cannot show what is involved in executing step 2, all we have is the mere promise that some acceptance property will give us the meaning of a certain word. (It would be like having a recipe for making a cake which no-one has ever tried to follow—we would be wary of putting it on the menu).

Notice three ways this can fail: the account may fail to be explanatory (e.g., it may fail to distinguish sufficiently between meaning properties), the account may turn out to be strongly relational (in which case the putative justification by deflationism about truth is void), or it

may fail to be basic (i.e., it may involve semantic notions in the *explanans*). I examine three conceptions of acceptance properties for the word 'red' for the purposes of executing the second step of the explanatory strategy. The first one fails to be explanatory, the second is relational, the third is not basic.

IV. Three attempts to explain aboutness.

i) Implicit definitions.

The first suggestion I shall look at goes like this: to appreciate that 'red' has the acceptance property it has is to be committed to use 'red' according to a (ramsified) implicit definition of the property of being red.

There is a well-known attempt at using our various platitudes, or core beliefs, about colors to implicitly define the properties of being red, orange, green, blue etc.[14] For redness we could have platitudes such as, for example: that the property of being red causes objects to look red to normal perceivers under standard conditions, that the property of being red is more similar to the property of being orange than it is to the property of being yellow, and so on. And similarly for the other colors.

We can list all the platitudes for all the colors in one long conjunction. This can give us an implicit definition of the colors. We can now turn this into an explicit definition by ramsifying: list in property name form all occurrences of the terms we are interested in, substitute variables for them, and quantify existentially over them. Then we can say that the colors, if there are any, are the properties which stand in the sorts of relations to each other and other things as described in what remains of the long conjunction. Furthermore, the

individual colors can be defined by singling out the distinct variables one by one. Thus, for the properties of being red and orange:

> (1) the property of being red = the *x* such that ∃*y*∃*z*...objects have *x* iff they look clearly *x* to normal perceivers under standard conditions, and *x* is more similar to *y* than it is to *z*...&...and so on.
>
> (2) the property of being orange = the *y* such that ∃*z*∃*v*...objects have *y* iff they look clearly *y* to normal perceivers under standard conditions, and *y* is more similar to *z* than it is to *v*...&...and so on.[15]

The suggestion is then this: to appreciate that 'red' has the acceptance property it has is to actually use 'red' according to the ramsified implicit definition for the property of being red. To spell it out:

> [R1] 'red' has the property of meaning RED = S is committed to apply 'red' to surfaces that have the property *x* such that ∃*y*∃*z*...objects have *x* iff they look clearly *x* to normal perceivers under standard conditions, and *x* is more similar to *y* than it is to *z*...&...and so on.

We can now gain an understanding of the crucial second step of the explanatory strategy by saying:

> [R2] If S is [R1]-committed, then S can use the very word 'red' to characterize its meaning (by saying, e.g., "'red' means RED").

I do not think Horwich would endorse this suggestion because I am not sure he believes 'red' is implicitly definable (we will come to the reasons why in Section (ii) below). Nevertheless, I have reason to believe it is an important suggestion to examine for it gives a clear idea of *conditional commitment* to use a word in a certain way. It ought to be possible, as Horwich stresses (p. 45, 90-92, Ch. 6), to mean what is generally meant by a word without endorsing statements that go into providing its basic pattern of use. To use Horwich's example (p. 90-91), we ought to be able to mean PHLOGISTON by 'phlogiston' even though we are not committed to the phlogiston theory which implicitly defines the word's meaning. Instead we can be committed to use the word in certain basic statements, on the *condition* that those statements are true. Such conditional commitment is necessary and sufficient to constitute a meaning for the word (as evidenced in the case of the meaningful word 'phlogiston').

We can use ramsified implicit definitions to flesh out what conditional commitment is. Unconditional commitment is commitment to use the word '*f*' according to the theory represented by '#'. Thus we have the unconditional acceptance property:

(3) #*f*.

Conditional commitment is then commitment to use '*f*' according to #, on the condition # is true, that is, on the condition that there is a unique something which has the properties and stands in the relations detailed by the theory, and which is thus a proper deserver of the name '*f*'. This tells us that ramsified implicit definitions are perfectly suited to express wherein this conditional commitment consists, thus we have the conditional acceptance property:

(4) $\exists x(\#x) \rightarrow \#f$.

It is commitment to use words according to conditionals like (4) which is necessary and sufficient for constituting meaning. Commitment to deploy unconditional regularities like (3) is sufficient for constituting meaning only if they entail conditionals like (4). As Horwich says (without any qualification) 'the acceptance properties that are constitutive of a meaning are conditional' (p. 92).

I take the import of this discussion of conditional commitment to be global. It holds for all acceptance properties, including the one for 'red', that they can constitute a meaning only if S can be conditionally committed to use the word according to it. This is a reason why I have tried to construe the acceptance property for 'red' as given in [R1]. It is unclear to me how else we might construe the notion of a conditional acceptance property for 'red'.

Unfortunately, the suggestion does not work. For on this conception of acceptance properties there will be counterexamples to the use theory of meaning. That is, some words, which we know have different meanings, come out as having the same uses, and therefore, according to the use-theory, the same meanings.

To demonstrate this I adapt Michael Smith's result that certain so-called network analyses are susceptible to a particular kind of 'permutation' problem: platitudes or core beliefs that implicitly define each member of certain tight-knit families of terms are not sufficiently distinguishable, even though the terms express different properties (Smith 1994, 2.11). If a similar case can be made for the use theory under the above conception of acceptance properties as implicit definitions, then, contrary to how the use theory would have it, conditional acceptance properties for different words would not instigate different meanings for them.

Consider again the suggestion for the acceptance property for ‘red’:

> [R1] ‘red’ has the property of meaning RED = S is committed to apply ‘red’ to surfaces that have the property *x* such that ∃*y*∃*z*...objects have *x* iff they look clearly *x* to normal perceivers under standard conditions, and *x* is more similar to *y* than it is to *z*...&...and so on.

and the account of the second step in the explanatory strategy:

> [R2] If S is [R1]-committed, then S can use the very word ‘red’ to characterize its meaning (by saying, e.g., “‘red’ means RED”).

Now notice that the conditional acceptance property for ‘orange’ has S being committed to apply ‘orange’ to surfaces that have the *very same* property as the property that S is committed to apply ‘red’ to:

> [O1] ‘orange’ has the property of meaning ORANGE = S is committed to apply ‘orange’ to surfaces that have the property *y* such that ∃*z*∃*v*...objects have *y* iff they look clearly *y* to normal perceivers under standard conditions, and *y* is more similar to *z* than it is to *v*...&...and so on.

And, correspondingly, when we set out the second step, like this:

> [O2] If S is [O1]-committed, then S can use the very word ‘orange’ to characterize its meaning (by saying, e.g., “‘orange’ means ORANGE”),

it should be understood that the only difference between S's use of the word 'orange' and S's use of the word 'red' is that they are different words, everything else in their use is exactly the same. Hence, when S says "'red' means RED" and "'orange' means ORANGE", then he or she *means the same* by the second occurrence of 'red' and the second occurrence of 'orange'. This is because S will be committed to use the two words 'red' and 'orange' in similar ways, and according to the use theory those two words will therefore have the same meaning when we have S executing the second step of the explanatory strategy. But we know that 'red' means RED and 'orange' means ORANGE. Therefore, this conception of the acceptance properties posited by the use theory of meaning is not workable.

The upshot is this. On this conception of acceptance properties we can make sense of the notion of conditional commitment, but acceptance properties cease to be properly explanatory: similar acceptance properties end up impossibly having to constitute different meanings.

Notice here a structural similarity to a very well-known argument against dispositional use theories of meaning. The core of Kripke's (1982) rendering of Wittgensteinian meaning scepticism is that dispositional use theories cannot account for how different tokens of the *same* word type can be associated with the *same* meaning (Kripke 1982, Ch. 2). The present argument says that in some cases the use theory (under this conception) cannot account for how tokens of *different* word types can be associated with *different* meanings.

ii) the acceptance property for 'red' is relational and is not implicitly definable.

The second suggestion for how we should conceive of acceptance properties stems from an objection to the above idea that 'red' can be implicitly defined. The objection goes like this:

‘red’ is not one of the words whose meaning can be implicitly defined, for the acceptance property for ‘red’ is in fact relational.

We might think that the only words that are implicitly definable are those whose acceptance properties traffic only in what sentences are regarded true and which inferences are found primitively compelling (thus we might be able to implicitly define ‘and’ and ‘true’). Since the acceptance property for ‘red’ also traffics in the sort of worldly situation where ‘red’ is applied, it is not amenable to implicit definition.

If ‘red’ is not implicitly definable, then there will be no permutation problem for ‘red’. And we may conjecture that permutation problems only arise where the acceptance property relies to some extent on relational paradigm cases.

This might make us re-think how we should conceive of acceptance properties. Appreciating that ‘and’ has a certain acceptance property might indeed be like being committed to use ‘and’ according to an implicit definition, but appreciating that ‘red’ has a certain acceptance property involves being committed to something which cannot be implicitly defined because it relies on relations to instantiations of redness.

The first thing we should notice about this, still only negatively described, suggestion is that it leaves us with no clear notion of what it would be for someone to be conditionally committed to apply ‘red’ in certain circumstances. The advantage, remember, of taking all words to be implicitly definable was that it gives us a clear idea of what conditional acceptance properties are like. If we have not got available the notion of implicit definition, then it seems to me there is little to impose a difference between being conditionally committed and being unconditionally committed to use a word in a certain way.

I think the second thing to notice is even more serious. We have seen that if ‘red’ is implicitly definable, then there will be permutation problems. If there are permutation problems, then there will be no adequate explanation of aboutness: words which we know are about different things will come out as having identical meanings. But now we are told that there is no such problem because the acceptance property for ‘red’ is relational. That is to say, we can explain aboutness for ‘red’ *only if* the acceptance property for ‘red’ is relational.

Now it seems to me we have made no advance over the old inflationist, relational accounts of aboutness. The deflationists as well as the inflationists are committed to saying this:

> (a) we can explain why ‘red’ means RED only if ‘red’ bears certain substantial relations to instantiations of redness.

If insisting on relationality is the core of the response to the permutation problems, then we have a suggestion which is non-relational—and so deflationary—in name only. It is like saying: it is a necessary condition for explaining aboutness that acceptance properties are relational, but - psst - their relationality doesn’t matter for their explaining aboutness. Obviously, this should be cause for concern for the deflationist, and we are, I think, owed an explanation of the sense in which deflationism can be deflationist in spite of having to insist on the inflationist’s tool of trade.

Notice that this is not simply the objection that some acceptance properties are relational. Horwich should clearly be allowed to operate with relational acceptance properties (see p. 41, 85-90). The objection is rather that, in the context of the permutation problem, it becomes necessary for the deflationist to insist on relationality in order to explain aboutness.

iii) What is it like to appreciate that a word has a certain relational acceptance property?
I have thrown doubt on the feasibility of conceiving acceptance properties as implicit definitions. Although treating acceptance properties as implicit definitions would give us a clear idea of the notion of conditional commitment, it fails to make them adequately explanatory.

As we saw, the deflationist may insist that, since the acceptance property for some potentially problematic words (e.g., 'red') are in fact relational, those words are not implicitly definable and hence the permutation problem is a red herring. But, on the assumption that this move works, it is left uncertain whether the deflationist's theory remains non-relational in any significant deflationary sense.

We now get to the third and last conception of acceptance properties, and indeed the conception which may be closest to Horwich's initial suggestion (i.e. before matters are complicated by the notion of conditional commitment). For at this point the deflationist may respond by pointing out that even though the acceptance property for 'red' is in fact relational, and needs to be in order to explain aboutness, we do not explain aboutness by *reading off* from this relational property what 'red' is about. Rather, it is the second step in the two-step explanatory strategy which allows us to explain aboutness. Thus, we should put ourselves in a position to appreciate that 'red' has a certain relational acceptance property, i.e. we should put ourselves in a position to use 'red' according to this property. This is the reason why that second step is so crucial for semantic deflationism: without it the use theory turns itself into inflationary semantics.

This gives us the following suggestion for the case of ‘red’. Conceive of the relational acceptance property underlying use of ‘red’ as (roughly) the disposition to apply ‘red’ to an observed surface when and only when it is clearly red. We need now to ask whether, if this is what acceptance properties are like, we can discharge the second step of the explanatory strategy in a basic and explanatory way. My suspicion is that this time we will get explanations, but that they are not basic.

The sense in which the account is intended to be *basic* is this: the *explanans* must not involve any semantic notions. For example, if the *explanandum* is the meaning property ‘*x* means RED’, then the *explanans*, some acceptance property, should not involve any semantic notions (p. 5). We should not confuse this notion of ‘basic’ with this stronger requirement: the *explanans* should not involve any semantic notions *and* we should be able to read off from it what the word is about. We have agreed to dispense with the reading off requirement as a condition of adequacy on a reductive account of meaning, so obviously we cannot fault Horwich for not satisfying it.

I want to argue that, in spite of the weakened requirement on basic explanatoriness, we do not get basic explanations. We cannot discharge the second step of the explanatory strategy without letting the *explanans* involve semantic notions. In order to show this I want to remind the reader of a similar story which is widely believed not to be explanatorily basic.

We might, in a somewhat Augustinian frame of mind, think that we can explain aboutness in terms of ostensive definitions. Thus, for the word ‘red’, we would explain what that word is about by pointing to a clearly red surface while saying “red”. We might add that we do not here grasp what ‘red’ is about by reading off anything from the relational features of such a pointing event (or indeed any other associated event). Instead we put ourselves in a position

to apply “red” when pointing to red surfaces. We obtain aboutness by becoming, as it were, able red-pointers. The parallel to step two in Horwich’s explanatory strategy should be obvious.

But this is not a basic explanation of the aboutness of ‘red’. The only way we can grasp the idea of someone being an able red-pointer is by assuming that he or she already have intentional thoughts (such as the thought expressible by ‘this is *F*’) about the thing pointed to. There is no such thing as a felicitous pointing which involves no intentional thought on the part of the pointer for you cannot perform a pointing without having a thought about something, such that you want your pointing to be about it.[16] Therefore, the ostensive definition theory of aboutness involves semantic notions and so fails to be basic, even in the required weak sense.

That story is sufficiently similar to the acceptance property story to suggest that the latter also fails to be basic. We are told that aboutness comes into place by someone appreciating that ‘red’ has a certain acceptance property, i.e. by using ‘red’ accordingly. That is, we obtain aboutness by becoming, as it were, able appliers of ‘red’ to clearly red surfaces. But it seems to me that, as in the pointing case, there is no such thing as a felicitous application of ‘red’ which involves no intentional thought on the part of the ‘red’-applier for you cannot apply “red” without having a thought about something, such that you want “red” to apply to it.[17] Just as we would expect the able pointer to have an intentional thought expressible by ‘this is *F*’ in order to point felicitously, so we would expect the able ‘red’-applier to have an intentional thought expressible by ‘this is *F*’ in order to “red”-apply felicitously. (Here ‘*F*’ can be any predicate. The complaint is not that the ‘red’ applier needs to involve intentional thoughts about the property of redness, but rather that felicitous ‘red’-application must involve some kind of intentional thought).

This demonstrates the sense in which the proffered explanation of aboutness is not basic in the required way. Our only way of understanding what it is for S to execute the second step of the explanatory strategy comprises a story that relies on semantic notions.

We can put the core of this objection in another way. Take it that felicitous use of 'red' in "'red' means RED" is use governed by the basic acceptance property for 'red', i.e. the inclination to use it for all and only surfaces that are clearly red. *In*felicitous use of 'red' would then be something like, say, use explained by the desire to employ 'red' in ways that the speaker might think pleases his or her interlocutors, or, use such that when S says "'red' means RED" this may be explained by a basic acceptance property for the schema "'…' means__" such that S is disposed to accept instances of such schemas when and only when they are homophonic (e.g., when it says "'red' means RED"). Clearly, such infelicitous use of "'red' means RED" does not succeed in characterizing a meaning of 'red': such uses are simply not of the meaning-constituting kind.

The question then is what we can say to distinguish felicitous from infelicitous use. If we cannot make the distinction, then we will not know what it is for S to properly discharge the second step of the explanatory strategy, for there needs to be a distinction between deployment of the locution "'red' means RED" that can characterize meaning, and deployment that can't. My objection is that it seems the only thing that distinguishes felicitous use from infelicitous use is that the former relies on engaging S's capacity for intentional thought.

This problem for the deflationary use-theory stems from the rejection of the reading off requirement. If we cannot read off from a meaning-constituting property what meaning it

constitutes, or even *that* it is meaning-constituting, then it becomes difficult to exclude non-meaning-constituting contenders like the two infelicitous properties I just mentioned. The problem goes to the heart of the deflationary theory. If use properties constitute meaning properties, then facts about use explain facts about meaning. Horwich has rejected the notion that facts about meaning can be read off from the facts about use (i.e. he rejects that the use story entails the meaning story *a priori*). Instead there is an appeal to a species of knowing-how (we get to know meanings by knowing how to use them), but, as one might have suspected in advance, an appeal to a species of knowing-how can easily fall short of convincing us that the story in one basic vocabulary makes true the story in another distinct, more superficial vocabulary.

I will make three comments about this objection. Firstly, the objection is not that, in order to understand an explicit *statement* of the acceptance property, one needs to already have semantic notions in place. This may or may not be the case, but is irrelevant since acceptance properties are not sentences or bits of language. Rather, the point is that we cannot understand what it is for a subject's use of 'red' to have a certain acceptance property, such that aboutness falls into place when he or she deploys it, unless we endow the subject with prior unexplained semantic abilities. What does the damage is not the conception of relational acceptance properties *per se*, it is when we try to execute the second step of the explanatory strategy that the problem occurs.

Secondly, this is not to deny that acceptance properties can trade in dispositions, inclinations and tendencies. The complaint does not flatly contradict what we have already accepted for the sake of argument, viz. that the use theory is potentially reductive inasmuch as it relies on dispositions, inclinations and tendencies. Rather, the complaint is that (reading off being disallowed) we cannot appreciate the explanatory role of those dispositions, inclinations and

tendencies unless we involve semantic notions, and thus they cannot be constitutive of semantic notions.

Thirdly, this is not the objection that we need to involve semantic notions in order to *identify* acceptance properties. We dispensed with this worry already at the beginning of Section II by acknowledging that what S accepts may be inferred from what she says and what she does. But there is a difference between knowing what S accepts and knowing what it is for her to accept it, and the present objection focuses on the latter. The objection is that, once we have identified a prospective acceptance property on the basis of what S says and does, then we need to involve semantic notions in order to let that property explain aboutness in a deflationary manner. Crucially, the deflationist cannot attempt to tone down the reliance on the second step of the explanatory strategy because that would just turn semantic deflationism into a theory which is indistinguishable from inflationary semantics (as I established in (i) and (ii) above).

To sum up, this kind of account manages to be explanatory in the sense that it can match the right words with the right meanings. But it does so by, as it were, piggy-backing on other semantic notions, such as the having of a thought expressible by 'this is *F*'. Therefore it is not basic.

iv) linguistic meaning and mental content.

Here is an initial diagnosis of this failure: this kind of account fails to be basic because it ends up explaining linguistic meaning-properties in terms of mental content-properties. It would not be surprising if this is correct for it would be odd if we could explain linguistic meaning without involving what people have in mind when they speak. On this diagnosis there are two options for the deflationary theorist. Either he or she accepts that we must

presuppose mental content-properties, and that therefore semantic deflationism does not provide basic explanations. Or he or she must intend to provide, as Horwich does (p. 99), a deflationary explanation of the aboutness of mental content-bearers too.

The second option is not open to the deflationist about meaning because all the problems I have elaborated so far will reproduce at the more fundamental level of mental content. There is no in principle difference between giving an account of the aboutness of linguistic meaning-bearers and of mental content-bearers. For linguistic meaning we ask what constitutes a word's having the meaning-property is has (e.g., what constitutes the word 'red''s meaning RED?). For mental content we ask what constitutes a mental content-bearer's having the content-property is has (e.g., what constitutes the content-bearer [red]'s having the content RED?). Horwich would have to posit acceptance properties (i.e. belief properties) for mental content-bearers too, and discharge the second step in the explanatory strategy in the same manner as before. But of course this does not do anything to escape the conclusion that we can only understand what it is to put oneself in a position to deploy the content-bearer in a certain way by telling a story that involves other semantic notions.

v) aboutness cannot be explained by a use theory justified by deflationism about truth. Serious problems arise for the account of aboutness which is justified by deflationism about truth once we appreciate that the deflationist must adopt the two-step explanatory strategy in order to account for aboutness in a distinctly deflationary manner. By examining three different conceptions of acceptance properties we found that the use theory fails to be, respectively, explanatory, deflationary, and basic. Justifying the use theory by deflationism about truth is thus not sufficient to give us basic explanations of the aboutness of words (or mental content-bearers).

V. Concluding remarks.

I think that Horwich has made impressive progress in the theory of meaning by identifying the relationality and 'reading off' requirements as the culprits in many failed attempts at giving a reductive account of meaning properties. But, if I am right, then semantic deflationism is not going to be as useful in providing an alternative as Horwich argues. The deflationary use theory is not especially parallel to deflationism about truth, and, as justified by deflationism about truth, it fails to give a reductive explanation of aboutness.

In Section IV(iv) I identified two options for semantic deflationism. Either attempt fundamental explanations of meaning (i.e. of both linguistic meaning-properties and mental content-properties). Or confine the deflationary approach to linguistic meaning-properties and allow involvement of semantic notions in the *explanantia* for different words.

Taking the second option would of course be inimical to Horwich's project for he is concerned to provide fundamental explanations. But given the problems that face this project we ought to re-think the merits of the second option.

It seems to me a very interesting philosophical question how words come to possess linguistic meanings, and what role mental content-properties play in this process. Even if we do presuppose mental content-properties, we might still wish for substantial accounts of first-person authority about meaning, understanding, and transmission of information in communication. In this I agree with Paul Boghossian (though he is a non-reductionist Platonist about mental content):

> [A]nti-reductionism [about mental content] is not only consistent with, but positively invites, a theory of the relation between thought and language. How do public language symbols come to acquire meaning and what role does thought play in that process? (Boghossian 1989, p. 549).

I think that some of the insights of semantic deflationism can play a role in this research project because non-basic, possibly non-relational, acceptance properties seems to me well suited to yield a novel account of linguistic meaning-properties, even though the semantic properties of linguistic items to some extent will be derived from mental content-properties.[18] Within this more modest project much can be gained from implementing Horwich's two-step strategy and trying to make do without the reading off requirement.[19], [20]

References:

Boghossian, Paul. (1989) "The Rule-Following Considerations," *Mind* 98, 507-49.

Carnap, Rudolf. (1963) "Replies and Systematic Expositions," *The Philosophy of Rudolf Carnap*, P. A. Schlipp (ed.), Open Court.

Evans, Gareth. (1985) "Semantic Theory and Tacit Knowledge," *Collected Papers*, Oxford: Oxford University Press.

Field, Hartry. (1994) "Deflationist Views of Meaning and Content," *Mind* 103, 249-85.

Geach, Peter. (1957) *Mental Acts*, London: Routledge & Kegan Paul.

Harman, Gilbert. (1982) "Conceptual Role Semantics," *Notre Dame Journal of Formal Logic* 23, no. 2.

Horwich, Paul. (1990) *Truth*, Oxford: Blackwell.

——. (1998) *Meaning*, Oxford: Oxford University Press.

——. (1999) "The Minimalist Conception of Truth," in *Truth*, S. Blackburn and K. Simmons, Oxford: Oxford University Press, 239-63.

Gupta, Anil. (1999) “A Critique of Deflationism,” in *Truth*, S. Blackburn and K. Simmons, Oxford: Oxford University Press, 283-307.

Jackson, Frank. (1998) *From Metaphysics to Ethics*, Oxford: Oxford University Press.

Johnston, Mark. (1988) “The End of the Theory of Meaning,” *Mind and Language* 3, 28-42.

——. (1992) “How to Speak of the Colors,” *Philosophical Studies* 68, 221-263.

Kripke, Saul. (1982) *Wittgenstein on Rules and Private Language*, Oxford: Blackwell.

Lewis, David. (1970) “How to Define Theoretical Terms,” *Journal of Philosophy* 62, 427-46.

——. (1972a) “Psychological and Theoretical Identifications,” *Australasian Journal of Philosophy* 50, 249-58.

——. (1972b) “Languages and Language,” *Minnesota Studies in the Philosophy of Science* 7.

—— (1992) “Meaning without Use: Reply to Hawthorne,” *Australasian Journal of Philosophy* 70, 106-110.

Loewer, Barry. (1997) “A Guide to Naturalizing Semantics,” *A Companion to the Philosophy of Language*, Bob Hale & Crispin Wright (eds.), Oxford: Blackwell.

Miller, Alex. (1997) “Boghossian on Reductive Dispositionalism about Content: The Case Strengthened,” *Mind and Language* 12:1, 1-10.

Quine, W. v. O. (1975) “Verbal Dispositions,” in *Mind and Language*, Samuel Guttenplan (ed.), Oxford: Oxford University Press.

Ramsey, Frank. (1931) “Theories,” in *The Foundations of Mathematics*. London: Routledge and Kegan Paul.

Schiffer, Stephen. (1987) *Remnants of Meaning*, Cambridge, Mass.: MIT Press.

Smith, Michael. (1994) *The Moral Problem*, Oxford: Blackwell.

Stoljar, Daniel. (2001) “Physicalism,” *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/physicalism/

Wright, Crispin. (1986) “How can the Theory of Meaning be a Philosophical Project?,” *Mind and Language* 1:1, 31-44.

——. 1992. *Truth and Objectivity*, Harvard: Harvard University Press.

Notes

[1] Horwich 1998. All subsequent page references are to this work, unless otherwise noted. *Meaning* is published simultaneously with the second edition of, *Truth*, Horwich’s influential account of deflationism about truth (first edition is Horwich 1990).

[2] E.g., Kripke 1982. For an overview of attempts to naturalize semantics, see Loewer 1997.

[3] I employ Horwich’s convention of giving meanings in capital letters, e.g., ‘dog’ means DOG.

[4] See also Stoljar 2001, Sec. 13. For the general picture see Jackson 1998, Ch. 1-3.

[5] The example is Horwich’s, 1999, p. 241-2. A more cumbersome alternative would be to have a device for substitutional quantification.

[6] See Ch. 4, and Horwich 1990.

[7] For dissent see, e.g., Wright 1992, Ch. 1; Gupta 1999.

[8] Of course it is also part of classical deflationism about truth (though not quite of Horwich’s brand) that the property of truth is not a substantial property at all. I do not think Horwich intends his theory of meaning to be parallel to deflationism about truth in this sense (cf. p. 42). Also, in order to be more genuinely deflationist the theory should probably be committed to saying that two sentences can be synonymous without having anything substantial in common, but this is not part of Horwich’s theory at all.

[9] Cf. p. 113: ‘We might call such a view of meaning, “deflationary”, both because it is *parallel* to, and because it is *justified* by, the deflationary view of truth’. Cf. also p. 10-11, 42, 69-71, 108-110.

[10] Quine, of course, was met with the same kind of objection and resorted to a purely behaviorist notion of ‘surface assent’. Cf. e.g. Quine 1975, p. 91.

[11] Though, we should not underestimate the semantic resources needed to perform this task: even after we have identified the accepted sentences we need to make divisions among them, assign some to the core uses, others to the periphery, and perhaps say that some (e.g., dead metaphors) belong to a basic regularity for another word.

[12] What follows relates to considerations found in Evans 1985, p. 336f. See also Wright 1986 Sec. I. Its ancestor is Geach’s (1957, p. 8) argument against behaviorism. The general point is that, where the task is to match up behavior with underlying states, we quickly run into diversity problems. Very dissimilar behavior may be explained by similar beliefs, and very different beliefs may be manifested by similar behavior.

[13] Notice that here we may want to posit different basic regularities for the same style of behavior: sometimes Jo opens the fridge because she desires to defrost it, sometimes because she wants to chill her own head on hot summer days etc.

[14] See, e.g., Johnston 1992. For this particular approach see Lewis 1970, 1972a and Jackson 1998 (building on Ramsey 1931 and Carnap 1963). I follow the exposition in Smith 1994, §2.11. See also Miller 1997.

[15] Here I follow Smith 1994, §2.11.

[16] Strictly speaking you can, e.g., if you are blindfolded and are playing a game where you point out who is ‘it’ (and even in this case you presumably have some kind of intentional thought of a narrow sort). But these are manifestly not the kinds of pointings the defender of the ostensive definition theory can rely on, precisely because blind pointings require that there are someone else there to *read off* from the incidence what is being pointed at.

[17] Again, strictly speaking you can, e.g., if you are blindfolded and are playing a game where you name unknown things “red” (though even this case seems to involve intentional thought of some narrow sort). But this is manifestly not the kind of case which the

deflationist can rely on because it requires there to be someone there who can *read off* this relational use what your utterance of “red” is about.

[18] The kind of account I have in mind is expressed well in Harman’s two-stage functionalism: ‘1. The meanings of linguistic expressions are determined by the contents of the concepts and thoughts they can be *used* to express. 2. The contents of concepts and thoughts are determined by their functional role in a person’s psychology.’ Harman 1982, p. 242, my emphasis. We also find it in a Lewis-style theory, which has whole thoughts, not concepts, as basic. Cf. Lewis 1992, p. 106 and Lewis 1972b. See also Loewer 1997.

[19] The first task that faces us it to make sense of the possibility of conditional commitment without employing ramsified implicit definitions. The answer is to employ implicit definitions which are not basic, i.e. which leaves spelled out paradigm cases in the *definiens* (see also Smith 1994, 1997).

[20] A version of this paper was read at the Wittgenstein and the Deflationary Turn conference, Sydney Chateau Hotel, 1997, thanks to the audience there, especially Paul Horwich, and thanks to Daniel Stoljar, Richard Holton, and Philip Pettit for comments and discussions.