Forthcoming in *The Philosophical Quarterly*

Meaning in Speech and in Thought

Stephen Schiffer
New York University

If we think in a lingua mentis, questions about relations between linguistic meaning and propositional-attitude content become questions about relations between meaning in a public language (p-meaning) and meaning in a language of thought (t-meaning). Whether or not the neo-Gricean is correct that p-meaning can be defined in terms of t-meaning and then t-meaning defined in terms of the causal-functional roles of mentalese expressions, it's apt to seem obvious that separate accounts are needed of p-meaning and t-meaning, since p-meaning, unlike t-meaning, must be understood at least partly in terms of communication. Paul Horwich, however, claims that his "use theory of meaning" provides a uniform account of all meaning in terms of "acceptance properties" that, surprisingly, implicate nothing about use in communication. But it turns out that the details of his theory belie his claim about it.

I. Introduction

The "use theory of meaning" (UTM) Paul Horwich advanced in his 1998 book *Meaning*, and updated in his 2005 book *Reflections on Meaning*,[1] claims that:

---

[1] Oxford UP, 1998 and 2005, respectively. All quotations of Horwich are from *Reflections.*
[2] See S. Schiffer, 'Truth and the Theory of Content', in H. Parret and J. Bouveresse (eds.), *Meaning and Understanding* (Berlin: Walter de Gruyter, 1981), pp. 204-22.
[3] 'Horwich on Meaning', *The Philosophical Quarterly*, 50 (2000), pp. 527-36.
[4] *Public-language* semantic facts pertain either to semantic speech acts (roughly, speech acts that entail speaker-meaning) or to the semantic properties that linguistic expressions or other things (e.g. smoke signals and traffic lights) have by virtue of the way they, or

(a) For every word *w*, there is a unique non-intentional property that explains *w*'s overall deployment, where "by 'the overall deployment of *w*', [he has] in mind the multitude of facts of the following sort:

- that *S* accepts certain sentences containing *w* (or would counterfactually accept them in certain circumstances);
- that certain *w*-sentences (or their mentalese correlates) articulate *S*'s desires (i.e. appear in *S*'s 'want-box') in certain circumstances;
- that those of *S*'s decisions that are articulated using *w* (i.e. that are constituted by appearances in *S*'s 'decision-box' of the mental correlates of sentences containing *w*) have a characteristic behavioral import." (pp. 37-8)

(b) For some kind *K* of sentences and kind *C* of circumstances, the non-intentional property that explains *w*'s overall deployment is the acceptance property of being such that the idealized law governing *w*'s use is that *w*-sentences of kind *K* are regularly accepted in circumstances of kind *C*.

(c) Since the acceptance property can't be an intentional property, accepting a sentence, as it figures in (b), can't be an intentional notion. "A picturesque way of [explaining acceptance] is to say that *S* accepts a sentence just in case that sentence, or its mental correlate, is in *S*'s belief box" (p. 41). (The "belief box" is defined by the fact that for one to believe a proposition is just for there to be tokened in one's belief box a sentence which expresses that proposition in one's mentalese. I introduced the belief-box metaphor in 1981 as a mnemonically convenient stand-in for whatever functional property makes a state a belief.[2] The same gloss applies, *mutatis mutandis*, to the other propositional-attitudes "boxes.")

Here are restatements of two examples Horwich offers as first approximations to meaning-engendering acceptance properties:

> The word 'red' means what it does by virtue of having the acceptance property of being such that the idealized law

[2] See S. Schiffer, 'Truth and the Theory of Content', in H. Parret and J. Bouveresse (eds.), *Meaning and Understanding* (Berlin: Walter de Gruyter, 1981), pp. 204-22.

> governing its overall deployment is that there is a propensity of the sentence 'That is red' to be tokened in one's belief box in response to the sort of visual experience provoked by observing a clearly red surface.
>
> The word 'and' means what it does by virtue of having the acceptance property of being such that the idealized law governing its overall deployment is that (all else being equal) $\ulcorner \sigma$ and $\sigma \urcorner$ will be tokened in one's belief box when, and only when, σ and σ' are tokened there.

In a review article on Horwich's *Meaning*,[3] I raised an objection to that book's articulation of UTM, which objection may be put in the following way:

> UTM, as befits a theory inspired by Wittgenstein, is explicitly abut the meaning properties of *public*-language expressions; it's a theory about "our ordinary conception of the meanings of terms in common, public languages, such as English and Spanish" (p. 41, n. 20). Now, the fact that words in a spoken language mean what they do has *something* essentially to do with their use in *communicative behaviour*, with the *speech acts* speakers perform in uttering sentences containing those words. A necessary condition for our words having meaning is that those words partly determine the meanings of the sentences in which they occur, and a necessary condition of, say, 'Does she drive a red Lamborghini?' having the meaning it has is that in a successful literal utterance of it the speaker would be *referring* to a certain female and *asking* his hearer if that female drives a red Lamborghini. Surely these observations are the merest platitudes about meaning in a public language; yet it's impossible to see how UTM can be made to square with them. In no sense does, say, 'red' mean

[3] 'Horwich on Meaning', *The Philosophical Quarterly*, 50 (2000), pp. 527-36.

> what it does in spoken English by virtue of having the acceptance property of being such that the idealized law governing its overall deployment is that there is a propensity of the sentence 'That is red' to be tokened in one's belief box in response to the sort of visual experience provoked by observing a clearly red surface. No such propensity to be tokened in the belief box can possibly explain the use of 'red' that determines its meaning in one's public language, and in this regard it's striking that what Horwich is referring to by 'the overall deployment of a word' includes nothing about its use to perform speech acts. Perhaps it is not implausible that Horwichian acceptance properties determine the meanings of "words" in one's neural system of mental representation, but meaning in a *lingua mentis* is quite different from meaning in a public language, even if in some sense one thinks in the same language one speaks.

In *Reflections on Meaning*, Horwich's 2005 sequel to his 1998 book *Meaning*, he explicitly allows that if my objection to his 1998 articulation of UTM is sound, then it's also a sound objection to his 2005 articulation of the theory, and he goes on to deny that it is a sound objection, and to argue against what he takes to be the positive view of meaning my objection implies, which he labels "the two-stage theory of meaning." But the purpose of this essay isn't to rejoin that old debate with him; it's rather to suggest that there is *now* no issue for us to debate. This is because my just-displayed sound objection to the 1998 articulation of UTM doesn't apply to what *Reflections* reveals to be the version of UTM Horwich *really* holds, and this because the "two-stage" theory my objection implies is actually *entailed* by that version of UTM. At issue is how meaning is constituted in speech and in thought.

II. The Gricean Two-Stage Theory of Meaning

Horwich sees my objection as stemming from my acceptance of a theory that, he says, I share with Jerry Fodor, Brian Loar, Stephen Neale, and others. According to Horwich, philosophers of language whose camp I inhabit

> advocate a two-stage theory: first, an account of how the terms of a mental language come to mean what they do; and, second, a neo-Gricean account of how the meanings of someone's overt public language derive from those contents. (p. 9)

Well, that vague characterization isn't a very accurate characterization of the theory on which my objection relied, a theory that is better, albeit more rebarbatively, called a *two-separate-accounts* (TSA) theory of meaning and that needn't be developed in a "neo-Gricean" way. At the same time, it's very clear from what Horwich says against TSA theories that might appropriately be called "neo-Gricean" (as we'll see, even the Gricean programme is more accurately called a TSA, rather than a two-stage, theory) that he believes UTM to be incompatible with the TSA theory my objection actually implies. Let me explain what I mean by beginning with the Gricean programme that Horwich takes to be his target.

The two-stage theory of meaning to which Horwich alludes is the Grice-inspired programme Intention-Based Semantics (IBS) for reducing all intentional facts—i.e. all public-language semantic facts[4] and all propositional-attitude facts—to facts that are intrinsically statable in wholly non-intentional terms. Horwich calls the programme a two-*stage* theory because those who developed it attempted to reduce the intentional to the non-intentional via two stages of reduction, but once we understand what the theory is it will be obvious that it's simply about what can be explained in terms of what (and therefore better called a two-separate-accounts (TSA) theory), and not about the order in which one tries to provide those explanations. Still, I'll follow Horwich's lead in initially

[4] *Public-language* semantic facts pertain either to semantic speech acts (roughly, speech acts that entail speaker-meaning) or to the semantic properties that linguistic expressions or other things (e.g. smoke signals and traffic lights) have by virtue of the way they, or their components, are used in communication. Examples of semantic properties that aren't public language semantic properties are the truth-values of propositions, the contents of perceptions and propositional attitudes (at least according to certain views), and the various model-theoretic constructions used to define logical truth.

characterizing the theory in terms of the two-stages by which its proponents have in fact proceeded.

In the first stage the IBS theorist seeks to establish two things: first, that a certain notion of speaker-meaning is definable, without reference to any public-language semantic notions, in terms of acting with certain audience-directed intentions; and second, that expression-meaning and all other public-language semantic notions are definable, without reference to any public-language semantic notions, in terms that include the propositional attitudes that define speaker-meaning. In the second stage the theorist seeks to reduce propositional-attitude facts to facts intrinsically statable in physical or topic-neutral terms and without reference to any intentional notions. Here there is no particular physicalist, functionalist or causal-functionalist approach favored by all IBS theorists: Brian Loar and David Lewis were straightforward commonsense functionalists; Jerry Fodor has his causal asymmetrical-dependence line; Dennis Stampe, Fred Dretske, and Bob Stalnaker seek reductions using a notion of "indication" in a related causal-functional way.

So, the theory Horwich invidiously compares to his own theory is about the relation between public-language semantic properties, on the one hand, and propositional attitudes, on the other. Before encountering Horwich's work on meaning, I had supposed that no one could coherently deny that the public-language semantic properties of expressions can't be understood without reference to the propositional attitudes constitutive of speaker-meaning and the other illocutionary acts we perform when we speak, and I knew of no philosopher of language who would deny it. I thought the only place where disagreement was possible was over the question whether an account of propositional attitudes could be given that made no appeal to the public-language semantic properties of our spoken words, as IBS claims, or whether, as Donald Davidson and others claimed, public-language semantic properties and propositional attitudes were inter-locking notions neither of which could be understood independently of the other. What surprised me about Horwich's UTM was that his explicit statements of the theory seemed, incredibly, to imply that the meanings our public-language words are not even partly determined by the way we use those words in the performance of speech acts. Presently I'll question the accuracy with which Horwich's explicit statements of UTM

represent the theory he really accepts. It's time to say something about how talk of a "language of thought" affects the issues at hand.

Beginning in the mid-seventies, with the publication of Gilbert Harman's *Thought* and Jerry Fodor's *The Language of Thought*,[5] those who sought reductionist accounts of propositional attitudes were attracted to the idea that we think in a language-like neural system of mental representation, the "brain's language of synaptic interconnections and neural spikes."[6] The notion of a "language of thought" has always been extremely vague, and it's doubtful that talk of a language of thought commits one to anything more than some pretty obvious assumptions, assumptions acceptable to just about any theorist, about functional relations that obtain among the neural states that realize our beliefs, desires and other attitudes. For example, Jerry Fodor and Zenon Pylyshyn, two of the most invested language-of-thought theorists, have argued that it was consistent with their understanding of a language of thought that connectionism was the correct account of the neural structures in which the language of thought was implemented.[7] For most of us who invoked the model (metaphor?) of a language of thought, its appeal was that we thought that conceptualizing issues about propositional attitudes as issues about meaning in a language of thought would cast the old familiar issues in a light that made them more tractable. For example, instead of staring blankly at the question "With what non-intentional relations between agents and propositions can we identify the belief and desire relations?" it was thought that one could get a leg up on answering it by in the first instance defining *believing* and *desiring* in the style of

> *x* believes *p* iff $\exists\sigma$(σ means *p* in *x*'s mentalese & is tokened in *x*'s belief box)
>
> *x* desires *p* iff $\exists\sigma$(σ means *p* in *x*'s mentalese & is tokened in *x*'s desire box)

---

[5] G. Harman, *Thought* (Princeton: Princeton UP, 1973); J. Fodor, *The Language of Thought* (Cambridge, MA: MIT Press, 1975).
[6] D. Lewis, 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61 (1983), pp. 343-77.; p. 346.
[7] J. Fodor and Z. Pylyshyn, 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition*, 28 (1988), pp. 3-71.

(and likewise, *mutatis mutandis*, for whatever other propositional attitudes can't be defined in terms of more basic ones), thereby reducing the question of what it is to believe a proposition to what were thought to be two more tractable sub-questions, and then, in the second instance, answering those two sub-questions by, first, identifying the language-of-thought relation with a non-intentional relation R and, second, finding non-intentional functional relations C_b and C_d to replace the box metaphors, so that the displayed place-holding definitions may be replaced by the proper reductions

x believes p iff $\exists\sigma[R(x, \sigma, p)\ \&\ C_b(x,\sigma)]$

x desires p iff $\exists\sigma[R(x, \sigma, p)\ \&\ C_d(x,\sigma)]$

Of course, what makes appeal to languages of thought attractive to IBS theorists is the plausibility of the claim that meaning in mentalese isn't about propositional attitudes the thinker has when she intentionally produces mentalese sentences—after all, neural sentences are never intentionally produced—but has rather to do with the conceptual roles of mentalese expressions and their causal relations to external states of affairs. For example, it's not implausible to say, at least as a first shot, that a symbol '#' means conjunction in x's language of thought provided that, if the sentence $\ulcorner\sigma \# \sigma'\urcorner$ is in x's belief box, then so are the sentences σ and σ', and if the sentences σ and σ' are in x's belief box, then so is the sentence $\ulcorner\sigma \# \sigma'\urcorner$. As the example suggests, from the perspective of IBS, Horwich's acceptance properties appear well-suited, at least structurally, to reduce the meaning properties of expressions in one's language of thought.

Conceptualizing issues about mental representation in terms of a language of thought gains some needed clarity—especially as regards the issues I'm joining with Horwich—by borrowing David Lewis's ingenious way of formulating the question of what it is for an expression to have meaning for a person or population of persons.[8]

A language, Lewis reminded us, is an abstract object that may or may not be used by anyone. Esperanto is a language no one speaks, and it's a contingent property of French that the French or anyone else speak it: there are possible worlds in which, though

[8] See D. Lewis, 'Language and Languages', in *Philosophical Papers*, vol. I (Oxford UP, 1983), pp. 163-88.

French exists and stands ready for use, no one uses it. It's a *necessary* truth that the sentence 'La neige est blanche' means *in French* that snow is white; but it's a *contingent* truth that it means that *among the French* or *for Pierre*. It's the *contingent* meaning relations that philosophers are concerned with when they wonder how meaning depends on use. When we think of a language as an abstract object which may or may not be used, it's convenient to think of it as a function from expressions—sequences of sounds, marks, smoke signals, patterns of neural firings, or whatever—onto meanings. Thus construed, and pretending that each expression of a language has a unique meaning in that language, we may then say that, for any language L, expression e, and meaning m,

e means m in $L =_{df} (L(e) = m)$,

thus making the fact that e means m in L a use-independent definitional truth. The interesting question, however—the one philosophers struggle with—isn't the trivially answered question of what is it for e to mean m in L, but rather the question, or some refinement of it, of what it is for e to mean m for *a particular person or population of persons*.

It may seem that the Lewisian structure enables us straightway to say

(1) e means m for x iff $\exists L$(e means m in L & x uses L),

thus enabling us to reduce the question of what it is for e to mean m for x to the question of what it is for x to use a language. But if (1) is moving in the right direction, it's moving there too quickly. The problem is that (1) doesn't put any restriction on *how* x must use L in order for some expression to mean for x what it means in L, and some restriction is required. For example, even though x has no idea what any L expression means, x might use a computer that runs a program written in L, and in that case no expression need mean for x what it means in L.

So what are the relevant uses? We need be concerned with only these two:

- One might use L as a *public language*.
- One might use L as a *language of thought*, as one's neural system of mental representation.

So e can't *simply* mean m for x; e can only mean m for x by meaning m in a language x uses as a public language of interpersonal communication or in a language x uses as a private language of thought (or uses in some other appropriate way, but for present

purposes I'm pretending that the only two relevant uses are use as a public language and use as a language of thought). It will therefore be convenient to stipulate that

(2) *e p*-means *m* for *x* iff $\exists L$(*e* means *m* in *L* & *x* uses *L* as a public language)[9]

(3) *e t*-means *m* for *x* iff $\exists L$(*e* means *m* in *L* & *x* uses *L* as a language of thought)

and now we may replace the inadequate (1) with

(4) *e* means *m* for *x* iff *e p*-means *m* for *x* or *e t*-means *m* for *x*.

Equipped with (4), we've reduced the question of use-dependent meaning to the following two questions:

QP: What relation must hold between a person *x* and a language *L* in order for *x* to use *L* as a public language? Let's call this relation—whatever it turns out to be—the *public-language relation*.

QT: What relation must hold between a person *x* and a language *L* in order for *x* to use *L* as a language of thought? Let's call this relation—whatever it turns out to be—the *language-of-thought relation*.

(For simplicity of exposition, I'll pretend that each of us has one public language and one language of thought.) One effect of the Lewisian reconceptualization is that in seeking a reduction of propositional-attitude relations the theorist now begins, not with

x believes *p* iff $\exists\sigma$(σ means *p* in *x*'s mentalese & is tokened in *x*'s belief box)

but with

x believes *p* iff $\exists L,\sigma$(*L* is *x*'s language of thought & ($L(\sigma)$ = *p*) & σ is tokened in *x*'s belief box),

hoping to replace that with the proper reduction

[9] Strictly speaking, we should want this to cover the case where x knows Hungarian and thus *could* speak and write it fluently, but in fact never gets to use it. But I'll ignore this slight complication in order to keep the present discussion as simple as possible.

x believes p iff $\exists L,\sigma[T(L, x)$ & $(L(\sigma) = p)$ & $C_b(x,\sigma)]$,

where T is the non-intentionally specified language-of-thought relation and C_b the functional relation that makes a state a belief.

When cast in terms of the Lewisian framework, IBS makes two big claims. The first is that the public-language relation can be defined, without reference to any public-language semantic notions, in terms of propositional attitudes that include those that define speaker-meaning. How might this be done? Well, perhaps, as a first approximation, by saying something like

> L is the public language of x iff x belongs to a population in which it's the practice to perform acts of speaker-meaning* by uttering sentences of L when what one means* fits the meaning that the sentence one uttered has in L,

where 'speaker-meaning*' abbreviates the complex conditions of the IBS account of speaker-meaning, and where talk of "fit" alludes to the way in which, for example, the proposition that Jane is affable fits the meaning of 'She is affable'. And then the second big claim is, of course, that the language-of-thought relation is definable in wholly non-intentional terms, perhaps terms that realize some commonsense or scientific propositional-attitude theory. We noticed above that in principle Horwich's acceptance properties appear well-suited to reduce the *t*-meaning properties of expressions in one's language of thought. And if they can play that role, then they are also well-suited to serve the IBS theorist's need for a reductive account of the language-of-thought relation, perhaps in something like the following way. First the theorist would define the notion of a *compositional acceptance theory for a mentalese language* as, roughly speaking:

> *T* is a *compositional acceptance theory for a mentalese language M with respect to x* iff *T* is a finitely axiomatizable theory whose axioms assign an acceptance property to each word and primitive structure of *M*, and *T* has for each sentence σ of *M* a theorem that assigns to σ an acceptance property *A* such that (1) σ's having *A* is logically equivalent to the parts and structure of σ having the acceptance properties *T*'s axioms assign to them and (2)

σ's both having *A* and being tokened in *x*'s belief box

metaphysically entails *x*'s believing the proposition $M(\sigma)$.[10]

Then the IBS theorist could say:

M is *x*'s language of thought iff there is a correct

compositional acceptance theory for *M* with respect to *x*.

(I trust it's clear that nothing in IBS precludes the possibility of our giving an acceptable sense to the idea that we think in a neural version of our public language. That the same language was both one's public language and one's language of thought would simply mean that one stood in both the public-language relation and the language-of-thought relation to the same language.)

Suppose now that we have correct IBS accounts of the public-language and language-of-thought relations, and therewith of *p*-meaning and *t*-meaning (see (2) and (3) above). For the IBS theorist, the public-language relation effects a reduction of public-language semantic properties to propositional-attitude properties and the language-of-thought relation effects a reduction of propositional-attitude properties to non-intentional physical or topic-neutral properties. Thus, according to IBS, while the language-of-thought relation can't be defined even partly in terms of the public-language relation, the public-language relation is easily definable in terms of the language-of-thought relation: all one has to do is to replace all talk of propositional attitudes in the IBS account of the public-language relation by the translation of that talk into talk about the *t*-meanings of mentalese sentences and the way those sentences may be tokened. Suppose we have that translation, and further suppose (for simplicity of exposition) that we have just one public language and just one language of thought, that these are unambiguous languages in which no two words *p*- or *t*-mean the same thing, and that for every word in a person's public language there is a word in her language of thought that *t*-means for her what the public-language word *p*-means for her, and vice versa. Then there will be a one-to-one

[10] Note that there being a base axiom that assigns an acceptance property to each word of *M* is consistent with the impossibility of specifying a word's acceptance property without referring to the acceptance properties of other words in the lexicon. As Horwich recognizes, that would just mean that the model for defining acceptance properties is the Ramsey-Lewis model for defining a theory's theoretical terms. See *Reflections on Meaning*, pp. 53-4.

function f, definable in wholly non-intentional (and no doubt partly causal) terms, from one's public-language lexicon onto one's language-of-thought lexicon such that, necessarily, for every phonetic word v of one's public language, v *p*-means for one what $f(v)$ *t*-means for one, and for every neural word n of one's language of thought, n *t*-means for one what $f^{-1}(n)$ *p*-means for one. Furthermore, if v *p*-means m for x, then v *p*-means m for x *by virtue of* the fact that $f(v)$ *t*-means m for x. For that is a straightforward corollary of the fact that, according to IBS, *p*-meaning facts are determined by the propositional-attitude facts to which they reduce, and those propositional-attitude facts are in turn determined by the language-of-thought facts to which they reduce. At the same time, if n *t*-means m for x, that n *won't t*-mean m for x *by virtue of* the fact that $f^{-1}(n)$ *p*-means m for x. For that it won't is a straightforward corollary of the fact that, according to IBS, since propositional-attitude facts aren't determined by public-language semantic facts, *t*-meaning facts are not determined by *p*-meaning facts. In this way, the IBS theorist will hold that, while a word's *p*-meaning derives from the *t*-meaning of its mental correlate, the *t*-meaning of a word is conceptually prior to, and therefore doesn't derive from, the *p*-meaning of its public correlate.

III. My Objection and the TSA Theory of Meaning It Presupposes

I consider the Grice-inspired programme of IBS to be hopeless; that has been my view since the mid-eighties.[11] There is no reason to think that we can give necessary and sufficient conditions for speaker-meaning wholly in terms of a speaker's non-semantic intentions, and it can be *shown* that the sort of account IBS requires of the public-language relation is unobtainable.[12] In his statement of UTM in his 1998 book, *Meaning*, Horwich claimed that his "acceptance properties" were the only meaning-constituting properties *any* word could have, including words in one's public language. My objection to that statement of UTM in my review article on *Meaning* was that whatever property of a word engendered its *p*-meaning, there could be no correct account of that property that

[11] My change of mind is documented in S. Schiffer, *Remnants of Meaning* (Cambridge, MA: MIT Press, 1987).

[12] See *Remnants of* Meaning, Chapter 9, and S. Schiffer, 'Is Gricean Semantics Defensible? Response to Anita Avramides and Stephen Neale', forthcoming in G. Ostertag (ed.), *Meanings and Other Things: Essays on Stephen Schiffer.*

didn't advert to the use of words to perform acts of speaker-meaning, and that therefore the *p*-meaning-engendering property of a word can't a Horwichian acceptance property. *Strictly speaking*, that objection doesn't *logically* entail a TSA theory of meaning, since it's not *logically* incompatible with the public-language relation also being the language-of-thought relation—but of course that is a preposterous idea: it's conceptually necessary that the public-language and language-of-thought relations be two distinct relations. The TSA theory implied by my objection, and which logically entails it, is just the following three claims, whose conjunction I'll call the *weak-two-separate-accounts theory* (WTSA) of meaning, 'weak' indicating that, as its less committed than IBS, it's the weaker of the two TSA theories.

(i) A word *w* can't *simply* mean *m* for *x*. If we assume that the only presently relevant uses of a language are as a public language or as a language of thought, then *w* can mean *m* for *x* by, and only by, either *p*-meaning *m* for *x* or else *t*-meaning *m* for *x*.

(ii) Since the public-language and language-of-thought relations are distinct relations, *p*-meaning and *t*-meaning are distinct notions, each requiring its own account (whence my commitment to a TSA theory of meaning).

(iii) There can be no explication of the public-language relation, or therefore of *p*-meaning, that doesn't advert to the use of that language to perform acts of speaker-meaning. In other words, a conceptually necessary condition for *L*'s being *x*'s public language is that *x* performs acts of speaker-meaning by uttering sentences of *L*. (In this regard, one would do well to keep in mind that talk of a "language of thought" is a fairly new way of talking, no older than a person who hasn't even reached middle age, and it's not entirely clear what sort of literal sense can be made of it. When philosophers, including Wittgenstein (Horwich claims that UTM derives from Wittgenstein), have wondered about how the use of a word

> determines its meaning, their sole concern was how the public-language use of the word determines its *p*-meaning. Certainly there can be no complete account of word meaning that doesn't tell us how use of a word determines its *p*-meaning.)

As I've already said, Horwich makes very clear that, in his view, if my objection to his 1998 articulation of UTM is sound, then it's also to a sound objection to the 2005 version of the theory. He then of course goes on to argue that it's not a sound objection to UTM, either as articulated in 1998 or as articulated in 2005. His response is to be found in his Appendix to Chapter 2 ("A Use Theory of Meaning") of *Reflections*, and to earlier parts of the book which the appendix aims to elaborate. The Appendix begins:

> The purpose of this appendix is to set out the reasons … for rejecting a two-level picture of meaning—a picture in which *mental* terms somehow mean what they do (i.e. embody the concepts they do), perhaps in the way described by UTM; whilst *public* terms derive their meanings, à la Grice, from the concepts that it is intended and agreed they express. (p. 57)

Now, as I have already made clear, my objection doesn't imply the IBS programme Horwich has in mind, so it's not surprising that some of the objections he levels against IBS are consistent with my objection and the TSA theory that sustains it, viz. (i)-(iii) just above. For example, he says that one reason to be sceptical of the Gricean programme

> is the sheer implausibility of supposing that our everyday literal use of familiar words is backed by Gricean intentions. The idea reeks of over-intellectualization." (p. 58)

Fair enough, but my objection is consistent with communication's consisting of intentions that could be possessed even by people who don't have Oxford degrees. At the same time, the arguments he marshals in support of his claim that his theory is a unary theory are intended to show that there can be no correct TSA theory, from which it would

follow that at least one of (i)-(iii) must be false. Witness, for example, these passages from *Reflections on Meaning*:

> [A]s we shall see in the appendix to Chapter 2, [it's arguable] … that [the] common meaning of [a phonetic word and its mental correlate] derives from the joint possession of the same meaning-constituting property …. [I]t is best to suppose that there is a *single* way in which meaning is constituted, applying equally well to both mental and overt languages. *Such an approach would obviously have to be non-Gricean* [my emphasis]. And it would be especially compelling if each of us thinks largely in our own *public* language. (p. 9)

> Instead [of a TSA approach to meaning], I favour a *uniform* account, which will deal in the same way with both overt and mental terms. (p. 31)

> To see … that the relation between a person's mental terms and their verbal expression is non-intentional … consider …. (p. 58)

Very well; but which of WTSA's claims (i)-(iii) would Horwich deny? Surely, he can't deny (iii). He can't possibly want to deny that to use a language as a public language requires communicating in it, that 'Where's your sister?' wouldn't have the *p*-meaning it has for us if we couldn't use it to ask someone where his sister was. If, however, he does accept (iii), then, since he holds that *t*-meaning is constituted by acceptance properties that implicate nothing about their bearers being used in communication, he must be committed to the public-language and language-of-thought relations being two distinct relations, thereby committing him, notwithstanding his protestations to the contrary, to a TSA theory of meaning, and thus to accepting (ii) along with (iii). Nor is it any easier to see how Horwich can deny (i). Clearly, it's necessarily the case that if *w p*- or *t*-means *m* for *x*, then *w* means *m* for *x*, and, since for present

purposes we're harmlessly assuming that use as a public language and use as a language of thought are the only relevant uses of language, then it's just as clear that, necessarily, if *w* means *m* for *x*, then *w p*- or *t*-means *m* for *x*, in which case *w* can mean *m* for *x* by, and only by, either *p*-meaning *m* for *x* or else *t*-meaning *m* for *x*, which is exactly what (i) says. So, to repeat, which of (i), (ii) or (iii) might Horwich have it in mind to deny?

The answer turns out to be *none of them.* Horwich's theory is entirely consistent with the IBS TSA theory, and it actually *commits* him to *accepting* (i)-(iii), the TSA theory that sustains my objection to his 1998 articulation of UTM. For my sound objection to his 1998 articulation is no objection at all to what his 2005 book reveals to be the use theory of meaning that he actually holds—again, notwithstanding his claims to the contrary. Let's turn to why that is so, and what might be leading Horwich to suppose otherwise.

IV. Horwich's TSA Theory of Meaning

The "short crude statement" (p. 28) of UTM that Horwich offers in *Reflections* at the beginning of Chapter 2 ("A Use Theory of Meaning") and then proceeds to elaborate for the rest of the chapter, is that:

> The meaning of a word, *w*, is engendered by the non-semantic feature of *w* that explains *w*'s overall deployment. And this will be an acceptance-property of the following form:—'that such-and-such *w*-sentences are regularly accepted in such-and-such circumstances' is the idealized law governing *w*'s use (by the relevant 'experts', given certain meanings attached to various other words). (p. 28)

(Horwich goes on to say that for a property φ to *engender* the meaning of *w* is for *w* to mean what it does *by virtue of* its having φ. I explained what he means by a word's "overall deployment" and by "accepting" a sentence at the beginning of this essay.) Now, the displayed statement of UTM certainly invites the surprised comment, "*Surely*, the *p*-meaning that, say, the pronoun 'you' has for me isn't engendered by any "acceptance property" but is instead engendered by a property the possession of which enables me to use 'you' to refer to the person to whom I'm speaking. Besides, 'you'

might not have any acceptance property, since my language of thought might not be a neural version of my public language." It turns out that Horwich really agrees with both points, that is to say, he agrees that using a language as a public language requires its use to perform intentional speech acts; and he agrees not only that he should allow for the possibility that one's language of thought isn't the natural language one speaks, and he even agrees that if in some suitable sense we think in the same language we speak, it doesn't follow that the acceptance property that engenders a word's *t*-meaning also engenders its *p*-meaning.

I'll presently consider Horwich's recognition that meaning in a public language necessitates the use of language in the performance of intentional speech acts, but let's first consider how he accommodates the relation between the neural language of thought and the phonetic public language. For that, let's return to the above quoted sentence fragment

> To see … that the relation between a person's mental terms and their verbal expression is non-intentional … consider …,

that I quoted above, only now let's put it in its complete context:

> To see … that the relation between a person's mental terms and their verbal expression is non-intentional … consider … (for simplicity) a person *S* who speaks and understands just one language, and suppose that none of its words is ambiguous. In that case there will be a certain causal correlation between these words and the terms of *S*'s language of thought. More specifically, … there is a one-one correspondence, *f*, between *S*'s public word-types, *v*, and his mental term-types, *n* [= *f*(*v*)] such that, whenever a sound sequence includes a token of *v*, the mental sequence that *S* uses to specify what was said includes a token of *f*(*v*). Conversely, if *S* has decided to say a certain thing and articulates what he has to say using a mental sentence containing a term, *n*, this will bring it about that *S* utters a

sequence of sounds containing $f^{-1}(n)$. (p. 58; I've substituted '*v*' and '*n*' for Horwich's '*w*' and '*m*')

What this suggests is that:

(A) For some causal correlation *f* intrinsically specifiable in non-intentional terms,

a phonetic word *v* *p*-means *m* for *x* by virtue of there being a neural word *n* such that (1) $f(v) = n$ and (2) *n* *t*-means *m* for *x*.

Let Φ be the acceptance-property by virtue of which *n* *t*-means *m* for *x*. According to (A), Φ wouldn't be the property by virtue of which *v* *p*-means *m* for *x*; rather, the property by virtue of which *v* *p*-means *m* for *x* would be the property of

(B) *being a y such that, for some z,* $f(y) = z \ \& \ \Phi(z)$.

And if it's *v*'s having (B) that determines it to *p*-mean *m* for *x*, then, while we can't say that the acceptance property Φ is *v*'s *p*-meaning engendering property, we can say that it's *p*-meaning is engendered by its being correlated in a certain way—viz. the way required (B)—with a neural term that has Φ, and that, consequently, that *v*'s *p*-meaning *derives* from the acceptance property of the neural term that is *v*'s image under *f*. Given the way *t*-meanings are related to propositional-attitude contents, Horwich would then be in complete agreement with the IBS theorist's claim that *p*-meanings derive from the contents of the propositional-attitudes with which they are related in a certain way.

So it should be obvious that (A) is compatible with both the IBS TSA theory and the less committed WTSA theory, but (A) doesn't entail WTSA, for it doesn't entail that theory's claim (iii), the claim that in order for a word to have a *p*-meaning for a person, she must be able to use the word to perform acts of speaker-meaning. But (A) isn't Horwich's final version of UTM. That is made clear in the following way, which returns us to Horwich's recognition that for a language to be one's public language, one must use that language to perform acts of speaker-meaning.

Horwich explicitly allows that a Gricean account of public-language meaning might be true! He says that not only doesn't he "deny that intentions play an important role in communication (p. 59)," but that UTM leaves him free to accept the Gricean's

claim that *u* means *that p* in community *C* just in case "there is an implicitly respected convention … within *C* to the effect that a speaker utters *u* only when he believes *that p*, and wants his audience to recognize that he does, and … etc." (p. 60). If, however, that Gricean account is correct, then, letting '$\Sigma(v)$' = 'sentence Σ containing word *v*' and '$Q(m)$' = 'proposition Q containing word-meaning *m*', it would seem that (A) would yield to

(C) For some causal correlation *f* intrinsically specifiable in non-intentional terms,

a phonetic word *v p*-means *m* for *x* by virtue of there being a neural word *n* such that

1) $f(v) = n$;
2) *n t*-means *m* for *x*; and
3) [1) & 2)] $\Rightarrow$ for every sentence $\Sigma(v)$ of *x*'s public language there is a proposition $Q(m)$ such that *x* belongs to a community in which there prevails a convention *C* to the effect that a speaker utters $\Sigma(v)$ only when he believes $Q(m)$, and wants his audience to recognize that he does, and … etc.

That, however, isn't very elegant, for the stuff in it about the correlation *f* is superfluous given that Horwich's claim about a Gricean account of expression-meaning already commits him to

(D) *v p*-means *m* for *x* just in case *x* belongs to a community in which there prevails a convention *C* wherein for every sentence $\Sigma(v)$ there is a proposition $Q(m)$ such that a member of the community conforms to *C* in uttering $\Sigma(v)$ just in case she believes $Q(m)$ and wants her audience to recognize that she does, and … etc.

In order to have a definition that makes explicit how a word's *p*-meaning derives from the *t*-meaning of its image under *f*, Horwich would need to replace all propositional-attitude talk with its equivalent language-of-thought talk, and then revise (A) to yield something roughly along the lines of

(E) For some causal correlation *f* intrinsically specifiable in non-intentional terms,

a phonetic word *v* *p*-means *m* for *x* by virtue of there being a neural word *n* such that

1) $f(v) = n$;
2) *n* *t*-means *m* for *x*; and
3) [1) & 2)] $\Rightarrow$ *x* belongs to a community in which there prevails a convention conformity to which requires that *x* not utter a sentence $\Sigma(v)$ unless there is a mentalese sentence $\Theta(n)$ and a proposition $Q(m)$ such that (a) $\Theta(n)$ *t*-means $Q(m)$ for *x*, (b) $(f(\Sigma(v)) = \Theta(n))$ & (c) $\Theta(n)$ is tokened in *x*'s belief box and the image under *f* of *x*'s phonetic sentence 'Let my hearer recognize that I believe $Q(m)$' is tokened in *x*'s intention box.

Now (E) is clearly compatible with IBS and, since Horwich is committed to some version of 3) that entails whatever propositional attitudes the public-language relation requires, he is committed to accepting WTSA. The puzzle is, how could he have thought his UTM was in opposition to IBS and every other TSA theory of meaning?

The answer, I believe, is supposed to be revealed in what he wrote immediately following his claim that UTM is compatible with (D):

> Now, from the perspective of UTM, the *correctness* of some such principle [as (D)] is not at all objectionable. What is objectionable, however, is the idea that it tells us *what it is* for an utterance to have a certain meaning. We should, instead, take the correlation articulated by (D) to be the product of two more fundamental facts: one to the effect—very roughly speaking—that members of a linguistic community tend to produce a sentence only when they intend to manifest their believing the proposition the sentence means … and the other to the effect that *believing* a given proposition is nothing more than *accepting* some [mentalese] sentence that expresses it …. Now [those two more fundamental facts] together entail (G). Moreover, neither of [them] takes a stand on what meaning *is*. So we see that … [(D)], though true enough, can be reconciled with any account whatsoever of [what engenders] facts of the form '*u* means *that p*' …; so it gives absolutely no information about the nature of meaning. (pp. 60-61)[13]

I find this puzzling. For one thing, how could (a) it be that the fact *that x belongs to a community in which there prevails a convention conformity to which requires one not to utter σ unless in doing so one means q* entails *that σ p-means q for x* but yet (b) not be the case *that σ p-means q for x* by virtue of the fact *that x belongs to a community in which there prevails a convention conformity to which requires one not to utter σ unless in doing so one means q*? It's perfectly consistent with the fact that a sentence *p*-means what it does by virtue of being governed by a certain convention that the convention-fact is in turn engendered by a causal-functional fact expressible in non-intentional terms

[13] My 'can be reconciled with any account whatsoever of [what engenders] facts of the form "*u* means *that p*" …' replaces Horwich's 'can be reconciled with any account whatsoever of how facts of the form "*u* means *that p*" are constituted'. For Horwich, to know how meaning-facts are constituted is just to know the what engenders them, i.e. what it is by virtue of which the those facts obtain. I made the change to bring the present wording in line with the statement of UTM (quoted above on p. 000) that he works with throughout Chapter 2.

about the workings of mentalese sentences in the heads of those who follow the convention. Also, while Horwich can say that the *p*-meaning of a phonetic word is engendered by the fact that it's causally correlated in such-and-such way with a neural word that has such-and-such acceptance-property, he can't say that it's engendered by the phonetic word's having the acceptance property, even if the neural and phonetic words are different versions of the same word. Furthermore, since a word can't *p*-mean anything unless it's used (or at least can be used) in intentional acts of communication, in whatever non-intentional way one specifies the property that engenders a word's *p*-meaning, that property must entail the convention-property whose possession by the word is entailed the word's *p*-meaning what it does. So the fact that Horwich's causal correlation *f* is non-intentional by virtue of being intrinsically specifiable in non-intentional terms is perfectly compatible with *f*'s being, or realizing, a propositional-attitude relation, in which case it's entirely compatible with a word's *p*-meaning being engendered by a non-intentional property that it's also engendered by a propositional-attitude property. And for still another thing, whatever Horwich means by 'engender', 'by virtue of', or 'constitutes', they are *his* terms, not terms used in the formulation of IBS. So no matter what he might mean by those terms, why should he think UTM is inconsistent with IBS? (At one point he remarks that "the intimate correlation between a public sound and a language-of-thought term is normally fixed during early childhood" (p. 58), and at that time the child's use of the public sound isn't backed by the complex intentions constitutive of speaker-meaning. For example, when the child acquires the word 'doggy' a little before his first birthday, he manifests that acquisition by pointing at every dog (and perhaps the odd fox, raccoon or Shetland pony) it sees and shouting 'doggy!'. But I don't see how Horwich can think anything of interest follows from that, since at the time the child first acquires 'doggy' he can't produce a sentence containing the word, nor in any other way use it to perform an act of asserting, asking, or requesting. So in whatever sense 'doggy' might express the child's first concept of a dog, the word at that time doesn't ***p***-mean anything for him.)

My guess is that Horwich was understandably struck by, but also unfortunately distracted by, the fact that we can account for a word's *t*-meaning in terms of an acceptance property it has, and that we can then go on to account for the use in speech of

a word in terms of the word's being causally correlated in a certain way with the use of a word—perhaps even the same word—in thought. This important observation does indeed entitle one to say that the *p*-meaning of a word is owed to, or derives from, the *t*-meaning of the same word, if the natural language that is one's public language is also one's language of thought, or, if the languages are different, that the *p*-meaning of a word is owed to, or derives from, the *t*-meaning of the mentalese word with which the public word is relevantly causally correlated. But that the public-language meanings of our words is inherited from the contents of the propositional attitudes we use those words to express has always been the Gricean's banner slogan, and is consistent with the TSA theory presupposed by my objection to Horwich's 1998 articulation of UTM.