Linguistic Content and Explanatory Psychological Content

Citation: Roberts, Gavin James (2013) Linguistic Content and Explanatory Psychological Content. [Thesis] (Unpublished)

Linguistic Content and Explanatory Psychological Content

Gavin James Roberts

Department of Philosophy

Birkbeck College, University of London

Submitted for the degree of MPhil Stud

Philosophy, 1 September 2012

Declaration

the work presented in the thesis is the candidates own

Gavin Roberts

Abstract

Linguistic Content and Explanatory Psychological Content

Burge (1979) presents an argument to show that externalism is true for mental content that incorporates the notion of a social term that is incompletely understood ('Social Externalism'). Burge relies on something like the following:

S→M Principle *We mean what we say*

Burge recognises that we do not always apply the S→M Principle. If one could identify some reasonably clear demarcation criteria (the Conditions) that could be applied to determine when the S→M principle should be upheld, we could formulate a substantial and potentially interesting positive thesis that is in fact stronger than Social Externalism (and thus entails Social Externalism). Such a thesis is the focus of this dissertation:

S→M Thesis: Interpreters are correct to apply S→M without qualification in (all) cases in which speakers misunderstand the social terms that they use, provided the Conditions are met.

In objection to Burge's position many writers have noted that upholding the S→M Principle in many cases results in belief ascriptions that fail to explain behavioural dispositions that the speaker has that are only explicable in light of the misunderstanding.

Since the Conditions determine when the S→M Principle holds it is hoped that the Conditions may enable us to accommodate intuitions on both sides of the debate.

Linguistic Content (as used here) is the state-of-affairs that the speaker actually represents by virtue of uttering the words in the context (determined, in part, by social facts). Explanatory Psychological Content (as used here) is the state-of-affairs that the speaker intends to represent by virtue of uttering the words in the context. When the S→M Principle holds Explanatory Psychological Content and Linguistic Content will coincide. When the S→M Principle does not hold, they will come apart.

The central theme that emerges is the trade-off between psychological sensitivity and semantic stability.

Acknowledgements

I would like to thank the many philosophers at Birkbeck College that have been so generous with their time and thoughts over the course of my studies, and most especially Professor Jen Hornsby, whose support was valuable beyond measure.

Above all though I thank my wife, Sam, without whom none of this would have been possible, and who also served as a regular reminder that there is a fine line between the profound and the banal…

Table of Contents

Page

OVERVIEW

PART 1
AN EXAMINATION OF BURGE'S SOCIAL EXTERNALISM

PART 2
EVANS, CRANE AND THE META-BELIEFS APPROACH

PART 3
PSYCHOLOGICAL SENSITIVITY VS SEMANTIC STABILITY: THE BROADER CONTEXT

OVERVIEW

(i) Introduction

This discussion is intended to be a friendly one amongst externalists; positions that would readily be described as semantic externalism and psychological externalism respectively are either assumed true or upheld.

I focus primarily on Burge's (1979) argument for psychological externalism (or rather what he calls 'anti-individualism'). Burge focuses on what we can call social terms – terms that have a public meaning that would *traditionally* be considered to be determined purely by social convention (such as 'arthritis', 'sofa' and 'contract'). Burge presents an argument to show that externalism is true for mental content that incorporates the notion of a social term that is incompletely understood and concludes that, in some cases at least, "differences in mental content are attributable to differences in the social environment" (p. 79). I will call Burge's position 'Social Externalism'.

Burge relies on something like the following:

S→M Principle *We mean what we say*

Of course what one means by this depends on how one is using the words 'mean' and 'say'. Setting this aside for now, the claim comes in two strengths:

S→M(always) *We always mean what we say*

S→M(sometimes) *We sometimes mean what we say*

Most people – indeed Burge himself - would accept that we do not always mean what we say in the following sense:

> **Extract A**
>
> If a generally competent and reasonable speaker thinks that 'orangutan' applies to a fruit drink, we would be reluctant, and it would unquestionably be misleading, to take his words as revealing that he thinks he has been drinking orangutans for breakfast for the last few weeks. Such total misunderstanding often *seems* to block literalistic mental content attribution…(1979, pp. 90/91, original emphasis).

Thus Burge himself recognises that there are “numerous situations in which we normally reinterpret or discount a person’s words in deciding what he thinks” (p. 89). It seems that Burge would agree that in this case (i) the speaker probably didn’t mean what he or she said (ii) there is a reasonable reading of what this person *said* that involves an orangutan (the ‘literalistic mental content’) (iii) in most contexts most would agree that it is unlikely that the person *meant* to make an assertion about an orangutan, and (iv) the reason for this discrepancy is that the speaker used a word that he or she did not understand; the speaker did not know that the word ‘orangutan’ does not refer to a fruit drink. Burge recognises that in cases like this (‘reinterpretation cases’) we reinterpret what a person has literally said (the literalistic mental content) in determining what that person meant to say and so he does not uphold S→M(always).

However, as suggested by the last sentence of Extract A, Burge does not accept that examples like this, which he says involve “quite radical misunderstandings” (p. 90) count against his claim. He emphasises that his conclusions depend “only on there being some cases in which a person’s incomplete understanding does not force reinterpretation of his expressions in describing his mental contents” (p. 92). In other words his central argument requires only that there be some situations in which

(i) a speaker uses a term that he or she misunderstands (or does not completely understand); and

(ii) we would uphold S→M in that situation.

He thus insists that despite such reinterpretation cases, “it is common practice, and correct, simply to take [the speaker] at his word” (p. 116). So Burge supports S→M(sometimes). In fact he is making two positive claims here:

1. It is common practice to apply S→M (to take people at their word); and
2. It is *correct* in these cases to apply S→M (to take people at their word)

Burge is certainly right about (1) and accommodating this fact is a difficult and important exercise. It is the second claim that is crucial.

If I am right, construed in the way in which the S→M principle is applied in Burge, the same principle serves in McGinn’s (1989) very quick argument from Putnam’s (1975) semantic externalism[1] to psychological externalism (which McGinn describes as the principle that the concept expressed by a term is given by what it means (1989, p. 31)).

Indeed the S→M principle seems to be quite commonly applied in externalist literature and one can see why: if one presupposes something like semantic externalism then what a speaker *says* is sometimes determined by factors that are external to the speaker. Applying the S→M Principle the natural conclusion to draw is that sometimes what we mean (and thus presumably what we think) is also determined by factors that are external to the speaker.

In objection to Burge's position many writers[2] have noted that upholding the S→M Principle in many cases results in belief ascriptions that fail to explain behavioural dispositions that the speaker has that are only explicable in light of the misunderstanding. As suggested by the earlier quote, Burge's response is not to deny that this is true in some (or even many) cases, but rather to point out that the argument needs to be made for *all* cases: for the objection to go through against his central conclusion it needs to be shown that there are *no cases* in which (i) and (ii) above hold. Although this defensive strategy offers a high degree of immunity from such objections, it leaves one questioning how substantial the thesis of Social Externalism really is. In other words, although we might agree with Burge that he has called attention to a "philosophically neglected fact about social practice" (p. 116): namely, that our attributions of mental content "do not require that the subject always correctly or fully understand the content of his attitudes" (ibid), it remains unclear when this philosophically neglected fact applies.

If one could identify some reasonably clear demarcation criteria (the Conditions) that could be applied to determine when the S→M principle should be upheld, we could formulate a substantial and potentially interesting positive thesis that is in fact stronger than Social Externalism (and thus entails Social Externalism). Such a thesis is the focus of this dissertation:

S→M Thesis: Interpreters are correct to apply the S→M Principle without qualification in (all) cases in which speakers misunderstand the social terms that they use, provided the Conditions are met.

[1] encapsulated in Putnam's conclusion that meaning, at least in the case of the meaning of natural-kind terms, is not "in the head" (1975, p.227)

[2] c/f Loar (1988, p 570-572), Crane (1991, p18-22) and Patterson (1990, 313-331)

What the S→M Thesis actually amounts to depends on what Conditions are proposed (Burge makes some suggestions about such conditions, which I use as a starting point). Since I believe that there is substance in both Burge's argumentation and in that of those who have raised objections, it is to be hoped that the Conditions would go some way to enabling us to accommodate our intuitions on both sides (of course we may find that there are residual tensions that are irreconcilable).

Some of those that have objected to Burge have done so from the internalist perspective. There are two general strategies, arguments from causation[3] and arguments from behavioural explanation. I will examine only the latter type of argument here, where the general strategy is to argue that the semantic content (i.e. what is said) is not sufficiently psychologically sensitive to provide an adequate account of an individual's behaviour; accordingly this can't be the psychological content.

It is part of my ambition here to develop this type of objection divorced from any commitment to internalism (implying that one need not adopt internalism in order to accommodate such concerns). On this view, making belief ascriptions that are behaviourally illuminating in the relevant ways does not require treating the individual as a 'brain in a box'; what it requires is recognition that the individual has a particular (and limited) epistemic perspective on the world. The importance of epistemic perspective is a general theme that runs throughout this discussion and being sensitive, in appropriate ways, to an individual's particular epistemic perspective on the world emerges as a key condition that needs to be met in order for the S→M Principle to apply.

Although I said at the outset that psychological externalism is here assumed or upheld, the analysis to be presented suggests certain *limits* on what conclusions ought to be drawn from a specific type of argument for psychological externalism; the type of argument that Burge (and McGinn) present.

(ii) <u>Justified belief ascriptions vs correct belief ascriptions</u>

Earlier I identified two distinct claims that Burge (1979) makes:

[3] In simplified from these arguments tend to run something like this (based on Crane, 1991)

P1. Person A's belief state will cause intentional behaviour

P2. Only intrinsic states are causally efficacious states

C. The element of Person A's belief state that causes intentional behaviour must be intrinsic

1. It is common practice to apply S→M (to take people at their word); and
2. It is *correct* in these cases to apply S→M (to take people at their word)

Burge points repeatedly to common practice in defence of his claim, i.e. in defence of his assessment of what psychological explanation (or at least 'mentalistic attribution' (p. 115)) is. The evidence for (1) is overwhelming and this surely counts for something; any plausible response to Burge's argument must recognise this. One promising avenue is to distinguish between being *correct* in applying S→M and being *justified* in applying S→M. The overwhelming evidence for (1) could then be interpreted as evidence for the following claim:

S→M Justification: Interpreters are *justified* in applying S→M without qualification in cases in which speakers misunderstand the social terms that they use provided the Conditions are met

We can see the distinction between S→M Thesis and S→M Justification as paralleling the distinction between holding a true belief and being justified in believing something. Another way of putting this is that one might argue that for pragmatic purposes we do apply the S→M principle and we are justified in so doing but that it does not follow that we are strictly correct in so doing. For now, I merely raise the possibility, which I will return to discuss more fully later.

(iii) The costs and benefits of upholding the S→M Thesis

There are costs associated with both upholding and with rejecting the S→M Thesis. These are dependent on (i) the costs and benefits of upholding vs rejecting the S→M Principle together with (ii) the Conditions that determine when the principle ought to be applied.

One factor that will inform the weighing up of the respective costs and benefits is what one takes the central aim of mental content attribution to be; i.e. what such attributions are intended to account for, or explain. In this extract Burge identifies three alternatives and highlights the one that his thought experiments emphasise:

> **Extract B**
>
> What I want to stress is that to a fair degree, mentalistic attribution rests not on the subject's having mastered the contents of the attribution, and not on his having behavioural

> dispositions peculiarly relevant to those contents, but on his having a certain responsibility to communal conventions governing, and conceptions associated with, symbols that he is disposed to use. It is this feature that must be incorporated into an improved model of the mental. (p. 115)

Burge is suggesting that the decision as to whether it is correct to reinterpret or uphold the S→M principle in a particular case will depend on how one weighs the following factors in determining what mental content attribution is appropriate:

I. the speakers 'true' understanding (i.e. whatever notion the speaker has "mastered")
II. the speaker's behaviour (specifically behavioural dispositions that are "peculiarly relevant" to the misunderstanding)
III. the speaker's responsibility to communal conventions (conventions "governing" and "associated with" the symbols he or she uses)

Burge's argumentation relies heavily on the fact that understanding comes in degrees (i.e. that our attributions "do not require that the subject always correctly or fully understand the content of his attitudes"). According to him, this should lead us to conclude that it is generally correct to apply the S→M Principle in cases in which a speaker incompletely or incorrectly understands a social term. Burge is surely right about understanding coming in degrees. This gives us good reason to resist (I) as a central aim of mental content attribution.

Of course aims (II) and (III) need not be mutually exclusive. However, sometimes the misunderstanding will be 'peculiarly relevant' (in Burge's words above) to the individual's subsequent behaviour and accordingly an attribution that does not take account of that misunderstanding will not account for the individual's 'peculiarly relevant' behaviour. This is an acute source of tension in upholding the S→M Thesis and the central debate thus seems to come down to weighing (II) up against (III) in cases like this. Burge's defensive strategy relies on the claim that sometimes we will opt for (III). However, more needs to be said if one is to defend the S→M Thesis.

On *face value* (and before attending to the Conditions), we have the following costs and benefits of applying vs not applying the S→M Principle:

The benefits of applying the S→M Principle

If speaker X seeks to express a belief by uttering "a is F" and hearer Y attributes a belief to X that she would express as "X believes that a is F" then Y's belief attribution will be correct and Y will have a correct or true understanding of X's beliefs; specifically, Y will be able to correctly specify the conditions under with X's belief would be true. This holds even if X actually misunderstands the meaning of the term 'a'. In other words it provides an extremely direct account of how thoughts and language relate to one another and how we communicate truth-conditional content to one another. Our ability to use language to communicate thoughts is in a sense *guaranteed*; the guarantee is part of how mental content attribution works.

The costs of applying the S→M Principle

The costs of applying the S→M Principle are that any behavioural dispositions that X has that are peculiarly relevant to his misunderstanding of the term 'a' will not be explained by attributing the belief *that a is F* to him.

The benefits of not applying the S→M Principle

If one does not apply the S→M Principle then this leaves it open that the actual content of speaker X's belief might be *that b is F* (where 'b' is a term that denotes whatever notion X had 'in mind' and which he wrongly thought was denoted by 'a') even though this is not the content that Y would ordinarily attribute to him. On this view the actual content of Speaker X's belief would explain behavioural dispositions X has that are peculiarly relevant to his misunderstanding of the term 'a'.

The costs of not applying the S→M Principle

The costs of course are giving up the direct account of how thoughts and language relate to one another. In the example above, Y might not have any basis on which to attribute the belief that b is F to X, since Y could be unaware of X's misunderstanding: strictly she ought to reinterpret but since she is unaware of the misunderstanding she would not do so. Since such situations might be expected to arise quite frequently, this threatens to drive a wedge between thought and language that threatens our ability to communicate effectively at all.

The view that what we say and what we think can come apart comes in varying strengths. Assuming that what a speaker *says* has truth conditions, we can distinguish a higher cost strategy from a lower cost strategy:

Higher Cost: A speaker's mental content (e.g. what a speaker believes) will *never* have truth conditions, i.e. what a speaker *says* has genuine truth conditions but what a speaker *thinks* does not

Lower Cost: A speaker's mental content *normally*[4] has truth conditions but when the speaker misunderstands a term that is used to express a thought those truth conditions are not the same as the truth conditions that attach to what he or she actually said.

The 'higher cost' strategy is associated with the two factor theorists (of which Putnam is one) and tends to go hand-in-hand with the causal argument that I mentioned earlier. I will touch on such positions only briefly.

(iv) Refining the S→M Thesis

What the S→M Thesis really amounts to comes down to what Conditions are proposed. As I discuss in Part 1, one of the conditions that Burge seems to suggest is that the misunderstandings are not relevant in the communication context (this is Condition 3 – refer Section 1.3). It is interesting to find this in Burge since it suggests that *the only times that we* knowingly *don't reinterpret is when we judge reinterpretation to be irrelevant in the context*. If this is right it makes the resultant position that can be drawn out of Burges work (i.e. the S→M Thesis) considerably more subtle than it might at first appear.

In addition, when we examine what would count as relevant in the communication context, it turns out that relevance would seem to lie in being relevant to behavioural dispositions that are 'peculiarly relevant' to the misunderstanding. In other words the suggestion is that Condition 3 should be filled out as:

Condition 3': The misunderstandings are not relevant to the speaker's intentions and expected behaviour in the communication context

However, if this is right then Condition 3' reconciles the tension that Burge draws attention to between (II) and (III) above: the speaker's responsibility to communal conventions would only apply when the speaker's misunderstanding about those conventions is irrelevant to the speaker's intentions and expected behaviour in the communication context. I should

[4] I say normally here because it remains possible that on this view, sometimes what the speaker meant to say did not have any truth conditions whilst what the speaker actually said did have truth conditions

stress that although admitting Condition 3’ leaves Burge’s central anti-individualist claim intact, the way that Burge treats cases like the arthritis case suggests that he would resist Condition 3’. Much of this dissertation focuses on direct and indirect reasons for accepting or rejecting Condition 3’.

If one grants Condition 3’ the result is that upholding the S→M Thesis does not entail bearing the cost of applying the S→M Principle that was drawn out above and indeed some of the benefits of not applying the S→M Principle may be available provided an appropriate reinterpretation is available in such cases. However, the flip-side of this is that the S→M Thesis does not deliver all the advertised benefits of applying the S→M Principle either, or, to look at this the other way around, it is subject to some of the costs of not applying the S→M Principle, specifically giving up the direct account of how thoughts and language relate to one another.

What emerges as a key recurring theme is that once one recognises that word-meanings are determined by factors unknown to the users of those words, e.g. Kripke’s causal theory of names, Putnam’s indexical theory of the meaning of natural kinds or Burges socially determined meanings of social terms, one will inevitably be faced with a trade-off between psychological sensitivity and semantic stability, because each subject’s behaviour (and judgements) will be determined, in part, by their particular (limited) epistemic perspective on the world.

A residual concern that comes out is that whether or not Condition 3’ is met seems to be somewhat ad-hoc. Attempts to avoid this ad-hocness do not seem to be open to us. My conclusion is that either one must grant a degree of ad-hocness in the process of belief attributions or one must give up on mental content attributions being genuinely psychologically sensitive.

(v) Structure of discussion

The central problem that lies before us is how to accommodate partial understanding or misunderstanding into mental content ascriptions (e.g. belief ascriptions).

In Part 1 I describe Burge's position more fully, clarifying the argument he presents, identifying candidate Conditions and then examining his case studies in some detail. I draw some preliminary conclusions in favour of the S→M Thesis and Condition 3'.

In Part 2 I introduce Evans's distinction between using a term and understanding that term in such a use and briefly discuss his application of that distinction to proper names. This serves as a point of comparison for Crane's similar strategy in response to Burge's arthritis case study (what I call the 'meta-beliefs approach'). I also define the notions of Linguistic Content and Explanatory Psychological Content and discuss some important disanalogies between the proper name and social term analyses. I conclude that although the analysis seems to be directing us towards adopting a combination of the S→M Thesis and the meta-beliefs approach, there remain significant residual concerns with this approach, primarily relating to Condition 3' (concerns that can be traced back to the tension between psychological sensitivity and semantic stability).

In Part 3, drawing on Kripke's 'A puzzle about belief', I argue that these residual concerns are structural in nature and should not necessarily count against the S→M Thesis. I also briefly examine one response to this problem, Stalnaker's version of bi-modal semantics and draw some morals for the S→M Thesis. In closing I suggest that the tension between psychological sensitivity and semantic stability is a problem that philosophers have been grappling with at least since Frege introduced the notion of sense. I conclude that the S→M Thesis (combined with the meta-beliefs approach) warrants further research and refinement.

PART 1

AN EXAMINATION OF BURGE'S SOCIAL EXTERNALISM

1.1 – Burge's Thought Experiment

Burge frames his position as 'anti-individualism' where individualism is the following claim:

Individualism: no difference in mental content without a difference in narrow content

Burge believes that his argument shows that individualism is false. Burge describes three steps to his thought experiment:

Step 1:

> **Extract C**
>
> A given person [Alf from here onwards] has a large number of attitudes commonly attributed with content clauses containing 'arthritis' in oblique occurrence. For example, he thinks (correctly) that he has had arthritis for years, that his arthritis in his wrists and fingers is more painful than his arthritis in his ankles, that it is better to have arthritis than cancer of the liver [etc]…he has a wide range of such attitudes. In addition to these unsurprising attitudes, he thinks (falsely) that he has developed arthritis in his thigh.
>
> Generally competent in English, rational and intelligent, the patient reports to his doctor his fear that his arthritis is now lodged in his thigh. The doctor replies by telling him that this cannot be so, since arthritis is specifically an inflammation of joints. Any dictionary could have told him the same. The patient is surprised, but relinquishes his view and goes on to ask what might be wrong with his thigh (1979, p. 77)

Step 2:

We are to imagine a counterfactual situation in which everything is the same with Alf, but in which:

> **Extract D**
>
> …physicians, lexicographers, and informed laymen apply 'arthritis' not only to arthritis but to various other rheumatoid ailments. The standard use of the term is to be conceived to encompass the patient's actual misuse…The person might have had the same physical history and non-intentional mental phenomena while the word 'arthritis' was conventionally applied, and defined to apply, to various rheumatoid ailments, including the one in the person's thigh, as well as to arthritis (1979, p. 78)

Step 3:

Step three is interpretational:

> **Extract E**
>
> It is reasonable to suppose that: In the counterfactual situation, the patient lacks some – probably *all* – of the attitudes commonly attributed with content clauses containing 'arthritis' in oblique occurrence. He lacks the occurrent thoughts or beliefs that he has arthritis in his thigh, that he has had arthritis for years [etc]…It is hard to see how the patient could have picked up the notion of arthritis [in the counterfactual situation]...'Arthritis', in the counterfactual situation, differs both in dictionary definition and in extension from 'arthritis' as we use it…So the patient's counterfactual attitude contents differ from his actual ones (1979, pp. 78/79)

The conclusion according to Burge is that "the patients mental contents differ while his entire physical and non-intentional mental histories, considered in isolation from their social context, remain the same….The difference in his mental contents is attributable to differences in his social environment" (ibid).

In summary, we begin with a situation in the actual world in which a patient (Alf) misunderstands the meaning of the term 'arthritis' but is still attributed beliefs about arthritis – some true and some false. Then we are asked to imagine a situation in the counterfactual world where the social environment is altered such the term 'arthritis' means *tharthritis* (which captures Alf's misunderstanding about the meaning of 'arthritis'). Some reflection on TwinAlf's situation leads to the conclusion that he surely doesn't have any beliefs that are about arthritis (as Burge points out, where would he have got the notion from). If we accept that in the actual world Alf did have at least some beliefs about arthritis, it follows that Alf and TwinAlf have differing mental contents and that these differences are attributable to differences in their social environments.

As Burge points out, most accept steps 2 and 3 and it is indeed hard to resist those. The focus is thus on the first step. In that step Burge provides the following premises: "he thinks (correctly) that he has had arthritis for years, that his arthritis in his wrists and fingers is more painful than his arthritis in his ankles" and "he thinks (falsely) that he has developed arthritis in his thigh" (Extract C). Either of these claims would be sufficient to support Burge's conclusions – i.e. we do not need to show that Alf had a false belief that he had arthritis in his thigh – it is sufficient to show that he had a true belief that he had arthritis in his wrists and fingers for years (and a misunderstanding about the meaning of arthritis).

There is no doubt that many people (including the doctor) would have readily attributed such beliefs to Alf. One might insist that intuitively it is correct to say that Alf, for example, thinks (correctly) that he has had arthritis for years. However, given that this assertion is combined with the assertion that Alf does not understand the term 'arthritis' it would be nice to find an argument in support of this assertion, i.e. an argument in support of Step 1.

Here is one suggestion as to how such an argument might go:

The Step 1 Argument

P1 The public meaning of the term 'arthritis' is *arthritis*

P2 The public meaning of the word 'arthritis' is determined by facts that include social facts

C1 When Alf utters the words ''I have had arthritis for years' Alf says something about *arthritis* (the meaning of which is determined by facts that include social facts)

P3 Alf misunderstands the meaning of the term 'arthritis' (he thinks it means *tharthritis*)

P4 When the Conditions hold, Alf means what he says (the S→M Principle applies)

P5 The Conditions hold

C2 When Alf utters the words 'I have had arthritis for years' he means something about *arthritis* (the meaning of which is determined by facts that include social facts), i.e. he holds a belief about/has a thought about *arthritis*.

If Alf's thought or belief is true then Alf thinks (correctly) that he has had arthritis for years. P3 is strictly irrelevant to the argument flow above (which is Burge's whole point really), but I include it since it is necessary in order for Steps 2 and 3 of the broader thought experiment to go through.

The scope of this argument depends on what Conditions are taken to hold. Shortly I will turn to examine what Burge has to say about such Conditions.

Before we move on, since Burge does not present things in exactly this way (as in the Step 1 Argument) it must be a good question as to whether Burge would support (or even formulate) this argument. To examine this question we will need to say a bit more about P4, i.e. just what is meant by "we mean what we say".

1.2 Words in oblique occurrence and the S→M principle

Loar (1988/1991) identifies the following principle at the heart of Burge's argument:

Loar 1 Differences in *de dicto* or oblique ascription imply differences in psychological content (1991, p. 570).

If the '*de dicto* or oblique' ascription of Alf's belief is what he says and Alf's psychological content is what he means, then the claim is equivalent to:

> Differences in what you say imply differences in what you mean

In other words, you mean what you say (P4)

Some care is needed here though, because this way of presenting things seems to take the notion of a *de dicto* ascription and an oblique ascription as equivalent. However, Burge does not use the terms in this way. McKay and Nelson (2010) identify three different conceptions of the *de re*/*de dicto* distinction:

Syntactically *de re/de dicto*: a sentence is *syntactically de re* just in case it contains a pronoun or free variable within the scope of an opacity verb that is anaphoric on or bound by a singular term or quantifier outside the scope of that verb. Otherwise, it is *syntactically de dicto*

Semantically *de re/de dicto*: a sentence is *semantically de re* just in case it permits substitution of co-designating terms *salva veritate*. Otherwise it is *semantically de dicto*

Metaphysically *de re/de dicto*: An attribution is *metaphysically de re* with respect to an object *o* just in case it directly attributes a property to *o*

What is clear is that what Burge calls 'oblique occurrences' are occurrences of terms in sentences that are *semantically de dicto*: as an example of a word in an oblique occurrence, Burge offers an example with the term 'water': given the facts that water is H_2O and that

Bertrand thinks that water is not fit to drink, it does not follow that Bertrand thinks that H_2O is not fit to drink (p. 76). A term will have an oblique occurrence in an intensional (i.e. not extensional) context (I will use the terms 'oblique context' and 'intensional context' interchangeably). This is what Burge says by way of *why* words like 'water' sometimes feature in non-oblique occurrences in this way:

Extract F

> Roughly speaking, the reason why 'water' and 'H_2O' are not interchangeable in our report of Bertrand's thought is that 'water' plays a role in characterizing a different mental act or state from that which 'H_2O' would play a role in characterizing. In this context at least, thinking that water is not fit to drink is different from thinking that 'H_2O' is not fit to drink…Clearly oblique occurrences in mentalistic discourse have something to do with characterizing a person's epistemic perspective – how things seem to him, or in an informal sense, how they are represented to him…(1979, p. 76)

On the other hand, when Burge uses the de dicto/de re distinction it seems to be in the sense of a mental attitude being *metaphysically de dicto/de re*. He suggests, for example, that the clearest cases of *de re* attitudes involve non-obliquely occurring terms in content clauses. In other words, the clearest cases of *metaphysically de re* attitudes involve sentences that are *semantically de re*. He has this to say about metaphysically *de re* attitudes in the context of his thought experiments:

Extract G

> When we say that Bertrand thinks of some water that it would not slake his thirst (where 'water' occurs in purely non-oblique position) we attribute a *de re* belief to Bertrand. We assume that Bertrand has something like an indexical relation to the water…It is easy to interpret such cases by holding that the subject's mental states and contents…remain the same. The differences in the situations do not pertain in any fundamental way to the subject's mind or the nature of his mental content, but to how his mind or content is related to the world…But what I want to emphasize here is that it is inapplicable in the cases our thought experiment fixes upon…We can appeal to attitudes that would usually be regarded as paradigmatic cases of *de dicto*, non-indexical, *non-de re*, mental attitudes or events. The primary mistake in the contract example is one such …What is crucial to our argument is that the occurrence of 'arthritis' is oblique and contributes to a characterization of the subject's mental content…the term occurs obliquely in the relevant cases and serves in characterizing

> the dicta or contents of the subject's attitudes. The thought experiment exploits this fact. (pp. 86/87)

One conclusion that we can draw from all this is that by 'oblique' Burge means *semantically* de dicto and by 'de dicto' he means *metaphysically de dicto*. Another is that on his view, the reason his claims have force is not because the attitude is '*de dicto*' but because it is *oblique*: "that the occurrence of 'arthritis' is oblique and contributes to a characterization of the subject's mental content" (ibid).

Something else that the above extracts draw out is the fundamental tension within Burge's thought experiments:

I. On the one hand, oblique occurrences of terms in mentalistic discourse "have something to do with characterizing a person's epistemic perspective": how things seem to the person (or are represented to that person in an informal sense of represented) (1979, p. 76);

II. On the other hand, when we ascribe beliefs we often do not take account of misunderstandings the person has concerning the meaning of the term, misunderstandings which are a reflection of the person's (limited) epistemic perspective

Here we find the trade-off between psychological sensitivity and semantic stability. It seems that in Burgean cases we want conflicting things: we want to be psychologically sensitive but we also want the semantic stability of the socially determined meanings of our terms.

Turning back to **Loar 1**, given what has gone before it would be less misleading to rephrase this as:

Loar 2 Differences in oblique (or *semantically de dicto)* ascription imply differences in psychological content.

However, when we formulate things this way it seems to be true *by definition*. Consider the following extracts from Burge:

> Clearly oblique occurrences in mentalistic discourse have something to do with characterizing a person's epistemic perspective (Extract F)

> the difference affects standard cases of obliquely occurring, cognitive-content-conveying expressions in content clauses (p.87)
>
> the occurrence of 'arthritis' is oblique and contributes to a characterization of the subject's mental content (Extract G)
>
> the term occurs obliquely in the relevant cases and serves in characterizing the dicta or contents of the subject's attitudes (Extract G)

It seems that according to Burge an obliquely occurring expression *is* a cognitive content-conveying expression. We can see why Burge holds this view when we reflect on what makes a context intensional/oblique – i.e. what makes it the case that co-referring terms cannot be substituted *salva veritate* in that context. The natural answer is the one that Frege gave us for the test of distinctness of *Sinn.*

The Intuitive Criterion of Difference[5]

> If two sentences are such that it is possible for a competent speaker to reflectively and sincerely accept the one and not the other, then they have different *Sinne* (*because* they have different cognitive values)

Applying the Intuitive Criterion of Difference to the sentences:

> "Bertrand thinks that water is not fit to drink"; and
>
> "Bertrand thinks that H_2O is not fit to drink"

leads rather directly to the conclusion that in this context the terms 'water' and 'H_2O' have differing cognitive values in this context (we need not frame our conclusion in terms of *Sinn,* since Frege's notion of Sinn was somewhat metaphysically loaded). And it is pretty clear that the reason these expressions occur obliquely (are semantically *de dicto*) is *because* the sentences are cognitive content-conveying and the terms contribute, in an essential way, to that content.

If Alf *expresses the belief* that he has had arthritis for years or if I *ascribe the belief* that he has had arthritis for years to Alf then the context is intensional (and 'arthritis' in oblique position) by definition. This is why Burge's point is a compelling one – people use language to express beliefs (and to report the beliefs of others). As a result the words used are in

[5] The terminology is due to Evans (1982, p. 18) and the definition is consistent with his

oblique occurrence; the occurrence is oblique because it is characterising a belief (or some other mental content). If those words have their public meanings in the scenarios that Burge describes (scenarios involving misunderstandings) then Burge's argument is sound.

So our focus should be not on whether this is an oblique occurrence of 'arthritis' or whether a difference in oblique ascription implies a difference in psychological content, but rather *what the meaning of 'arthritis' is in that oblique occurrence.* Burge's answer is that it is the public meaning of arthritis. This is the assumption that needs examination. We could reformulate the argument above to draw this out more explicitly:

The Reformulated Step 1 Argument

C1 When Alf utters the words ''I have had arthritis for years' Alf says something about *arthritis* (the meaning of which is determined by facts that include social facts)

P4' The term 'arthritis' is being used to express or characterise Alf's belief, i.e. it is an oblique occurrence of 'arthritis'

P4'' If the Conditions hold then the term 'arthritis' has the public/linguistic meaning of 'arthritis' in this oblique occurrence

P5 The Conditions hold

C2' Alf believes that he has had arthritis for years (Alf means something about arthritis)

This is, I believe, the argument that Burge is asking us to accept if we grant Step 1. The crucial premise here is P4''. Once it has been observed that the term 'arthritis' is in an oblique context it is easy to overlook this additional premise that is required in order to derive the conclusion that Alf means what he says.

1.3 The necessary conditions for upholding the S→M Principle

As discussed at the outset, given that all agree that S→M does not always hold, it would be reasonable to demand some demarcation criteria that would enable us to separate cases in which the S→M Principle holds from those in which it does not. Burge makes the following general remarks about differences between cases in which reinterpretation is standard and when it is not:

> **Extract H**
>
> A person's overall linguistic competence, his allegiance and responsibility to communal standards, the degree, source, and type of misunderstanding, the purpose of the report - all affect the issue…For purposes of defending the thought experiment and the arguments I draw from it, I can afford to be flexible about exactly how to generalize about these various phenomena. The thought experiment depends only on there being some cases in which a person's incomplete understanding does not force reinterpretation of his expressions in describing his mental contents… [such cases] appear to be legion (pp. 91/92)

Towards the end of his paper he summarises things thus:

> **Extract I**
>
> The key feature of the examples…was the fact that we attribute beliefs and thoughts to people even where they incompletely understand contents of those very beliefs and thoughts…Crudely put, wherever the subject has attained a certain competence in large relevant parts of his language and has (implicitly) assumed a certain general commitment or responsibility to the communal conventions governing the language's symbols, the expressions the subject uses take on a certain inertia in determining attributions of mental content to him. In particular, the expressions the subject uses sometimes provide the content of his mental states or events even though he only partially understands, or even misunderstands, some of them. Global coherence and responsibility seem sometimes to override localized incompetence.
>
> The detailed conditions under which this "inertial force" is exerted are complicated and no doubt a little vague… (p. 114)

Burge goes on to identify one such necessary condition:

1. Clearly the subject must maintain a minimal internal linguistic and rational coherence and a broad similarity to others' use of the language (p. 114)

However, he notes that it is "hardly sufficient" and suggests that we should add an etiological consideration:

2. In cases in which the speaker developed his linguistic habits from others who had distinctively regional conventions, we take the person to be "committed to using the words according to the conventions maintained by those from whom he learned the words" (p. 114)

But he goes on to note that the situation is still more complicated than this since a person "might simply decide unilaterally" (p. 114) to follow some other usage or make up his own usage, thus "self-consciously opting out" (p. 114). In such a case Burge holds that members of his community should reinterpret him accordingly. He thus adds a third condition:

3. The individual's intentions or *attitudes toward communal conventions* and communal conceptions (which "seems more important than the causal antecedents of his transactions with a word", i.e. more important than the etiology). (p. 114)

For our purposes we can capture the key elements of the suggestions above in two conditions:

Condition 1: The subject must maintain a minimal internal linguistic and rational coherence and a broad similarity to others' use of the language

Condition 2: The subject is committed to using the words according to the conventions maintained by those from whom he learned the words

Burge is not particularly interested in setting out to find a detailed list of necessary and sufficient conditions of this type, since he believes this would not be "philosophically interesting"; on his view, what is interesting is the "philosophically neglected fact about social practice: Our attributions do not require that the subject always correctly or fully understand the content of his attitudes" (p. 116). However, to the extent that such conditions restrict the scope of the Step 1 argument, such conditions could be interesting. Burge includes what I take to be a crucial discussion a little later that seems to suggest another condition. Here are two relevant extracts:

Extract J

For almost any content except those that directly display the subject's incomplete understanding, *there will be many contexts in which it would be misleading to attribute that content to the subject without further comment*. Suppose I am advising you about your legal liabilities in a situation where you have entered into what may be an unwritten contract. You ask me what Al would think. It would be misleading for me to reply that Al would think that you do not have a contract (or even do not have any legal problems), if I know that Al thinks a contract must be based on a formal document. *Your evaluation of Al's thought would be crucially affected by his inadequate understanding.* In such cases, it is incumbent on us to cite the subject's eccentricity: "He would think that you do not have a contract, but then he thinks that there is no such thing as a verbally based contract." (p. 91, my emphasis)

Extract K

We do not ordinarily seek out true object-level attitude contents to attribute to victims of errors based on incomplete understanding. For example, when we find that a person has been involved in a misconception in examples like ours, we do not regularly reinterpret those ascriptions that involved the misunderstood term, *but were intuitively unaffected by the error*. An attribution to someone of a true belief that he is eating brisket, or that he has just signed a contract, or that Uncle Harry has paid off his mortgage, is not typically reformulated when it is learned that the subject had not fully understood what brisket (or a contract, or a mortgage) is. Moreover, we shall frequently see the subject as sharing beliefs with others who understand the relevant notions better. In counting beliefs as shared, we do not require, in every case, that the subject 'fully understand' the notions in those belief contents, or understand them in just the same way (pp. 93-94)

Burge makes two important qualifications here. In Extract K he suggests that one of the conditions that is required for us to apply the S→M principle is that the content ascription is "intuitively unaffected by the error". In Extract J he suggests that in certain contexts we should qualify our application of the S→M principle ("it would be misleading to attribute that content to the subject without further comment"). In the overall context this seems to amount to recognition that in these cases the belief attribution would be misleading without some qualification (which I will suggest amounts to reinterpretation[6]).

This suggests that Burge recognises that we should admit as one of the Conditions:

Condition 3: The misunderstandings are not relevant (in a way or ways to be defined further) in the communication context

In summary then we have three candidate conditions that Burge seems to propose in his paper:

Condition 1: The subject must maintain a minimal internal linguistic and rational coherence and a broad similarity to others' use of the language

Condition 2: The subject is committed to using the words according to the conventions maintained by those from whom he learned the words

Condition 3: The misunderstandings are not relevant (in a way or ways to be defined further) in the communication context

[6] Recognising that sometimes such reinterpretation is implicit (see Section 1.4)

1.4 Developing Burge’s case studies

The Contract example

Burge offers the example of a speaker who misunderstands the public meaning of the word ‘contract’ and thinks that one cannot have a contract with someone unless there is a written agreement when, in actual fact, no formal document is required for two people to enter into a contract.

Let’s imagine that Bill is such a person and that Bill’s promoter (Promoter 1) is aware of Bill’s misunderstanding. He is having a conversation with another promoter:

Promoter 1 (S1): “Bill agreed a contract with Jack under which Jack would pay him £1million if he sang on Wednesday night”

Promoter 2 (S2): “But Bill went to the Flamingo club and got paid £100,000 for singing the same songs on the same night”

Promoter 1 (S3): “Yes, right after he agreed a contract with Jack he signed a contract with the Flamingo club”

Promoter 2 (S4): “Well I’m not doing business with him again the man’s behaviour is completely unpredictable, not to mention illogical”

Promoter 1 (S5): “Actually he’s very reliable and logical it’s just that he thought there was no such thing as a verbally based contract, so he didn’t believe that he had agreed a contract with Jack.”

Promoter 2 (S6): “You mean that Bill doesn’t know what a contract is?”

Promoter 1 (S7): “No, he knows what a contract is, he just didn’t realise that a contract can be entered into verbally or in writing”

We find evidence both for and against the application of the S→M principle here. On the one hand the fact that Promoter 2 has been misled as evidenced by S4 suggests that some qualification of the notion of a contract would be relevant in the context. In this case the qualification follows in S5. However, interestingly, the qualification in this case does not seem to take away from the intuition that Bill still knows what a contract is (as in S7). Burge would no doubt cite this as evidence that the S→M principle applies in this case.

One point worth raising here is that there is more than one way of interpreting the evidence: one could interpret the latter fact as evidence that the necessary conditions for something being of the social kind *contract* do not include being potentially verbally-based (by 'social kind' I mean a non-natural kind[7]) . This might be the reason that we don't feel that Bill misunderstands the notion of a contract. One might argue that in this case Bill believed (correctly) that the (necessary) conditions for being a contract include, for example (i) its being binding (ii) its being made between at least two parties (iii) its requiring at least one of the parties to moderate behaviour in light of the contract etc. He also believed that (i.e. if you asked him he would agree that) it is not possible for a contract to be verbally-based. If it's being potentially verbally based is not a *necessary* condition of being a contract then Bill does not misunderstand the social term 'contract'.

However, I don't think that this is a very promising response since in this case it seems to be more-or-less irrelevant. This is because even if we accepted that being potentially verbally based was not a necessary condition of being a contract, we can still construct scenarios in which reinterpretation would seem to be appropriate in light of this particular misunderstanding. Assume once again that Bill's promoter (Promoter 1) is aware of his misunderstanding. We can then imagine the following dialogue:

Bill (to Promoter 1): "I'm not going to enter into a contract with Jack: I'm going to give Jack the impression that he has a deal but I'm not going to sign anything. I want to keep my options open about where I play tomorrow night" (from which Promoter 1 infers that Bill's misunderstanding about contracts is in play)

Promoter 2 (to Promoter 1): "Does Bill intend to enter into a contract with Jim?"

Let's assume that Promoter 1 knows that Promoter 2 has been advising Bill not to enter into a contract with Jack. If Promoter 1 answers with the following it would clearly be misleading:

Promoter 1 (to Promoter 2): "No he doesn't intend to enter into a contract with Jack" (Report 1)

[7] Although I suspect that social factors play a large part in determining the extension of many 'natural kinds'

It's misleading because Bill intends to behave in a way which is likely to result in him entering into a contract with Jack and this is just what Promoter 2 is advising him not to do.

The opposite report is of course equally misleading

Promoter 1 (to Promoter 2): "Yes he does intend to enter into a contract with Jack" (<u>Report 2</u>)

The only way of adequately explaining Bill's intentions is by taking account of his misunderstanding (even if it's a contract-related misunderstanding about a condition that is not a necessary condition for something being of the social kind *contract*).

We can imagine Promoter1 responding in something like the following way:

Promoter 1 (to Promoter 2): "No he doesn't intend to enter into a contract with Jim, but he may well do so in any event as he thinks that you have to sign an agreement to have a contract with someone" (<u>Report 3</u>); or

Promoter 1 (to Promoter 2) "Yes he does intend to enter into a contract with Jim but he doesn't realise it because he has a misunderstanding about contracts" (<u>Report 4</u>)

Since what seems to be required is a qualification of the *meaning* of the word 'contract' in the example above, this probably counts against the suggestion that being verbally based is not a necessary qualifying condition of the social kind contract. However, whether or not one draws this stronger conclusion it certainly shows that something beyond this distinction would be required to account for the linguistic evidence. What the linguistic evidence suggests is that the misunderstanding, whatever its nature, is relevant in the context.

It will be noted that this is very similar to Burge's treatment of the example in Extract J. Now presumably Burge would point to something like Report 3 as evidence in favour of the S→M principle being applicable here, since it seems to attribute a belief about contracts (the public concept) to Bill ("he thinks that you have to sign an agreement to have a contract with someone"). However, the belief attribution about 'contracts' is *qualified* here since the misunderstanding is made explicit. The reason for the qualification is that the misunderstanding is relevant in the context – what this suggests is that Condition 3 really is crucial to the S→M thesis.

The Sofa example

Burge suggests the following example of a misunderstanding concerning the public meaning of 'sofa': "In addition, he might think that sufficiently broad (but single-seat) overstuffed armchairs are sofas" (p. 80). This is what we might call an error of inclusion – he thinks the social kind is broader than it actually is, whereas the contract case is an error of exclusion.

Once again we could imagine this not being grounds for failure to properly understand what a sofa is. We can imagine saying the following:

> 'He understands what a sofa is, he just doesn't know that broad overstuffed armchairs are not sofas'

However, once again we can still construct scenarios in which reinterpretation would seem to be appropriate. We can imagine a situation in which Bob knows that Bill believes that broad single-seat overstuffed armchairs are sofas but did not find it necessary or appropriate to correct him at the time. We can then imagine the following dialogue:

Bill (to Bob): "I'm going to buy a sofa that fits in that nook" (pointing to a nook in the room that is not large enough to accommodate a sofa, from which Bob infers that Bill's misunderstanding concerning broad overstuffed armchairs being sofas is in play)

Jim (to Bob): "What's Bill going to buy today?"

If Bob answers with the following it would clearly be misleading (in fact it would be false):

Bob (to Jim): "He's going to buy a sofa" (<u>Report 1</u>)

We can imagine the subsequent conversation going as follows:

Jim (to Bob): "Why is he going to buy a sofa?"

Bob (to Jim): "Because he believes that a sofa could fit into the nook in his living room" (Report 2)

As evidence that it would be natural for Bob to reinterpret what Bill has said, it is clearly more natural for Bob to report Jim's intentions as follows:

Bob (to Jim): "He intends to buy an armchair" (Report 3)

He might even elaborate about Bill's misunderstanding in order to avoid confusion later (particularly, for example, if he knew that Bill had asked Jim to help him transport his purchase home and so might ask him for help moving a 'sofa'):

Bob (to Jim): "He's going to buy an armchair, but he thinks that broad overstuffed armchairs are sofas" (Report 4)

Once again we find that Condition 3 is crucial to the S→M thesis. We might ask then in what way is the misunderstanding relevant in the communication context? (This was left to be spelled out when Condition 3 was initially formulated.) It seems that the misunderstanding is taken to be relevant in the communication context when it would result in a *misrepresentation of the speaker's intentions and expected behaviour*. We can see that this is the concern over Reports 1 and 2. If, for example, Bob was going to buy a two-seater sofa to put in a large space in his living room, then using Report 1 to report Bob's state-of-mind would be unproblematic. When we attribute a state-of-mind to somebody we aim to account for their intentions and expected behaviour in the context – failure to do so is a misrepresentation of that person's state-of-mind.

Behind all of this is the point of central importance: that whether we reinterpret the public meaning of the word *depends on the context* AND *we only don't reinterpret when the misunderstanding* is *not relevant to the content in the context* (relevant in the sense of misrepresenting the speaker's intentions and expected behaviour).

Revisiting the arthritis example

In the arthritis case, as Burge sets it up, the misunderstanding does seem to be relevant to Alf's intentions and behaviour in the context.

Let's imagine that in the actual world Alf and his doctor have the following exchange:

Alf: “I am concerned that my arthritis has lodged in my thigh”

Doctor: “That’s not possible – arthritis is specifically a condition of the joints”

An exchange that would have gone something like follows in the counterfactual world:

TwinAlf: “I am concerned that my arthritis has lodged in my thigh”

TwinDoctor: “Yes, that’s a possibility, we should do some tests on that”

It seems as if the doctor has attributed the following belief to Alf in the actual world:

> Alf believes that *arthritis* is a condition of the joints and muscles

Once again, Burge would point to this as evidence that the term ‘arthritis’ is not reinterpreted in this case. However, we can imagine the doctor going on to make the following reports to another doctor:

Doctor: “Alf thinks that you can get arthritis in your muscles” (Report 1)

Doctor: “Alf does not really know what arthritis is – he thinks it’s a condition of the joints and muscles” (Report 2)

Doctor: “Alf has a misunderstanding about the meaning of ‘arthritis’ he thinks that ‘arthritis’ means a condition that you can get in your joints and your muscles” (Report 3)

The point being that I don’t see that any more, or less, information is provided to the second doctor under any of these reports of Alf’s beliefs (i.e. since the same knowledge would be gained from each for all intents and purposes these belief ascriptions amount to the same). Since Reports 2 and 3 are explicit about Alf’s misunderstanding about the concept arthritis and the meaning of the word ‘arthritis’ respectively, this suggests that Report 1 is providing the same information implicitly. In other words, this is a *qualified* application of the S→M principle and the reason it is qualified is because Condition 3 is not met (in this case the misunderstanding is relevant and would result in a *misrepresentation of the speaker’s intentions and expected behaviour* in the context).

If the examples above and the history of philosophical counterexamples is taken into account, it seems likely that for almost any misunderstanding over the meaning of a social term, it will be possible to construct situations or contexts in which reinterpretation would be correct or appropriate in light of this cashing out of Condition 3.

1.5 Preliminary Conclusions

If we assume for now that the analysis is along the right lines then it seems that the following are necessary conditions of applying the S→M principle:

The Conditions:

Condition 1: The subject must maintain a minimal internal linguistic and rational coherence and a broad similarity to others' use of the language

Condition 2: The subject is committed to using the words according to the conventions maintained by those from whom he learned the words

Condition 3': The misunderstandings are not relevant to the speaker's intentions and expected behaviour in the communication context

It is worth noting that it is not whether we *know* about the misunderstanding that is relevant, it is whether, in the context, the misunderstanding is relevant to the content. Of course if one was unaware of the misunderstanding then one would be ignorant that the misunderstanding was relevant to the content. This suggests that there are cases in which we may apply the S→M principle in error: specifically in cases when Condition 3' does not hold and yet as interpreters we don't realise this. It would be reasonable to conclude that in cases like this, in which the misunderstanding is relevant to the content but we are unaware of this fact, we may still be *justified* in applying the S→M principle but strictly we are *not correct* in so doing (strictly we have misrepresented the individual's mental state).

One positive proposal we could formulate would be a proposal under which interpreters are justified in applying the S→M principle (recall the discussion in the Introduction, part (ii)). Something like the following:

S→M Justification: Interpreters are *justified* in applying S→M without qualification in cases in which speakers misunderstand the terms that they use provided the Justification Conditions are met

Justification Conditions: The interpreter is unaware of the misunderstanding or the interpreter is aware of the understanding and has *no reason not to believe* that the Conditions hold. I think that this much is certainly supported by Burge's thought experiments and it supports Burge's claim (surely correct) that it is common practice to apply the S→M Principle.

However, what we are principally interested in is the S→M Thesis. I suggested at the outset that one of the reasons for investigating the Conditions was that it might enable us to accommodate intuitions on either side as to whether the S→M Principle should be applied or not.

In earlier discussion it emerged that the tension that Burge's thought experiments concerning misunderstandings focuses us on is that:

I. On the one hand, oblique occurrences of terms in mentalistic discourse "have something to do with characterizing a person's epistemic perspective" (1979, p. 76);

II. On the other hand, when we ascribe beliefs we often do not take account of misunderstandings the person has concerning the meaning of the term, misunderstandings which are a reflection of the person's (limited) epistemic perspective

Now we can see that the context-sensitivity of Condition 3' enables us to go part-way to reconciling this tension. When the misunderstanding that results from that person's particular epistemic perspective is relevant, in the communication context, to that person's intentions and behaviour, then we reinterpret the misunderstood term accordingly.

However, even though Burge himself was the source of Condition 3 I think that it would be significantly overstating the case to suggest that we have accommodated all of Burge's intuitions here. If we take the arthritis example discussed earlier, I think that it is reasonably clear from what he says in (1979) that he would resist the claim that the meaning of the word 'arthritis' is qualified when Alf says "I think my arthritis has spread to my thigh" (see pp. 77-79): on his interpretation of the thought experiment the term 'arthritis' just means *arthritis* in this sentence, and further, since it is an oblique occurrence he concludes that Alf holds a belief about arthritis.

If this is right then either he will need to resist the specific formulation of Condition 3 as Condition 3' or he will need to insist that Condition 3' has been met in this particular case. I suspect that he would press on the specific formulation of Condition 3': Recall that we have found Burge suggesting that the decision as to whether it is correct to reinterpret or uphold the S→M principle comes down to how one weighs the following factors in determining what mental content attribution is appropriate:

i. the speaker's behaviour (specifically behavioural dispositions that are "peculiarly relevant" (Extract B) to the misunderstanding); versus
ii. the speaker's responsibility to communal conventions (conventions "governing" and "associated with" the symbols he or she uses (Extract B))

Here is an important extract from Burge:

> **Extract L**
>
> It does not follow from the assumption that the subject thought that a word means something that it does not (or misapplies the word, or is disposed to misexplain its meaning) that the word cannot be used in literally describing his mental contents. It does not follow from the assumption that a person has in mind something that a word does not denote or express that the word cannot occur *obliquely* (and be interpreted literally) in that-clauses that provide some of his mental contents. (p. 101, my emphasis)

This drives to the root of Burge's position. His claim is that the notion of thoughts (specifically thought contents) that we make use of in our everyday interactions with one another is not concerned primarily with how things actually are with the individual that the thoughts are attributed to. Or, to put this slightly differently, on Burge's view, everyday psychological explanations are less concerned with a person's intentions and specific behavioural dispositions than has previously been supposed.

Burge takes it that his thought experiments show that (ii) should be weighed above (i). If this is right then presumably he would resist (i) being incorporated into a condition for the application of the S→M principle (e.g. into Condition 3). I hope that my argumentation above shows that this does not accord with our intuitions in the cases examined. It seems reasonably clear that it is precisely the peculiarly relevant behavioural dispositions that determine whether we think the misunderstanding is relevant in the context or not. In other words, it seems that we only weigh (ii) as more important when it does not conflict with (i).

If this is the case then it begins to look as if what Burge's thought experiments really focus us on is not so much a speaker's responsibility to communal conventions but rather the importance of context in our ascriptions of mental content (i.e. in determining what state-of-affairs has been mentally represented in a given circumstance – see later). In order for us to ascribe a belief about a social kind, we require an appropriate level of 'mastery' to support the discussion in which the speaker is engaged. What seems to be required is that the speaker must understand the term in the relevant respects. When the misunderstanding is not relevant we apply the S→M principle. It seems that our demands for what counts as understanding depend on the communication context. One might say it depends on what the speaker is *using* the word to convey.

Our commitment to communal conventions emerges more clearly in S→M Justification: it is because each of us *implicitly takes on the responsibility to use terms that we believe we understand sufficiently well in the context* that as a community we are justified in presupposing that people have a reasonable grasp of the words that they use. This would explain why for practical purposes we apply the S→M principle provided we have no reason to suspect that there is a relevant misunderstanding in the context.

On the evidence so far it seems that we ought to conclude that Burge should grant Condition 3'. What about holding that the condition is met in the arthritis case discussed earlier? Against this response we have the little argument presented at the end of Section 1.4 concerning the implicit qualification of the meaning of the term 'arthritis' (to the effect that no more, or less, information is provided to the second doctor under any of those belief reports).

One piece of evidence that Burge points to in defence of his position is that speakers admit error when their mistakes are pointed out to them, as when Alf in the arthritis case says something like 'Oh I see, well obviously I was wrong to believe that I had arthritis in my thigh, what do you think it could be?' rather than saying something like 'But doctor when I said 'arthritis' I meant [tharthritis]…' However, there are two responses available here:

Response 1: the first response is to make the rather obvious point that Alf may have altered the concept that he associated with the word 'arthritis' now that the misunderstanding has been pointed out to him.

Response 2: the second response involves a counter to an objection to the first response. The objection is that Alf will have misrepresented his own belief state if he refers to his prior belief as a belief about *arthritis* when he actually held a belief about *tharthritis*. The counter is that, once again, the

> use of ‘arthritis’ is qualified in the above sentence, so there is no inconsistency (and if the misunderstanding was irrelevant to one of his prior beliefs no qualification would be necessary)

One might begin to wonder on what grounds Burge’s resistance is to be based. One final source of resistance to Condition 3’ might be the overly high costs associated with granting this condition. As I discuss in Section 2.8 there are certainly issues that must give one pause for thought. However, how high the costs are depends in part on how plausible it is that we can find appropriate reinterpretations when Condition 3’ is not met. Crane’s meta-beliefs approach is one strategy to reinterpretation that I will examine in some detail in Part 2. Two points are worth making before we move on:

i. since Burge admits that reinterpretation is required in some cases, he is in just as much need of a theory that can accommodate the reinterpretation cases as his detractors (of course that does not mean that he needs to endorse the meta-beliefs approach)
ii. on face value some form of reconciliation with Burge’s *central claim* is still available if one adopts the S→M Thesis, i.e. it would follow from the fact that we do apply the S→M Principle in some cases in which we are aware of (or suspect) a person’s misunderstanding (or partial understanding) that Burge’s central claim (that *sometimes* differences in mental content are attributable to differences in the social environment) still holds

With respect to the second point above, we would find positively, for example, for Burge’s suggestion that what he calls ‘mastery’ of a concept is not required in order to attribute a mental attitude in which that concept features. However, although on face value we do apply the S→M Principle in some cases in which we are aware of a person’s misunderstanding, it begins to look as if the only time that we don’t reinterpret is when the misunderstanding is irrelevant to the speaker’s intentions and expected behaviour in the context.

This makes the resultant doctrine of Social Externalism more subtle than it might seem to be in less examined form.

PART 2

EVANS, CRANE AND THE META-BELIEFS APPROACH

2.1 Evans and the use/understanding distinction

The crucial point to bear in mind when examining arguments like Burge's (and McGinn's which I will touch on later) is that they *begin* with the assumption that words (or sentences) in a public language have a public or linguistic meaning that is determined, in part, by factors external to the speaker , i.e. they presuppose semantic externalism. Here is a definition of semantic externalism, due to (Lau and Deutsch, 2010, p. 4) which I will call Linguistic Externalism since it is concerned with the meanings of words:

Linguistic Externalism: the thesis that the (linguistic) meaning and reference of *some of the words* we use are not solely determined by the ideas we associate with them or by our internal physical state

I will assume that Linguistic Externalism holds. I recognise that I have not presented an argument for Linguistic Externalism here and I take myself to be addressing those that share this commitment. Evans, who seems to share this commitment has the following to say:

Extract M

> Once one's interest is in the phenomenon of language itself, one must be concerned with the way in which it functions as a means of communication among members of a community…One will then regard the utterances of individual speakers of the language as exploitations of a *linguistic system which exists independently of anyone's exploitation of it*[8]…There immediately opens up the possibility of a gap between what a speaker means to say…what thought he wishes to express…and what he strictly and literally says according to the conventional meanings of the words he utters (1982, p. 67, my emphasis).

If we accept that many words (and the sentences that contain them) have *public meanings that are independent of a* particular speaker's understanding thereof, what seems to follow as a matter of course is that sometimes a person can *use* a word without fully understanding it or even without understanding it at all - I will call this *the*

[8] I presume that he means of any individual's exploitation of it – that is that he should not be taken as suggesting that language might exist independently of the existence of people, i.e. of anyone *at all* exploiting it.

use/understanding distinction. Crane (1991) notes that the distinction itself should not be controversial since it is needed to make sense of ambiguity and punning, for example (p. 19).

If one is going to promote arguments of the form of Burge's that rely on the S→M Principle (despite the misunderstanding) then one will need to explain why there is no such gap in the cases in which the arguments are applied, i.e. one will need to defend a version of the S→M thesis. Of course there are other ways of coming at the problem of how language and thought relate to one another. However, *if one is going to come at it this way* (i.e. beginning with Linguistic Externalism), *then what requires defence is a version of the S→M thesis.*

In *The Varieties of Reference* Evans examines the relationship between referring terms and thoughts about the individuals referred to. Thus we find Evans's commitment, closely related to Extract M above, that "the notion of *using a term* to refer is a less fundamental notion than the notion of *understanding a term* in such a use" (p. 398, my emphasis). Evans's suggestion is that we should distinguish between:

- Using a term to refer; and
- Understanding a term in such a use

Which is associated with the difference between:

- referring to something; and
- thinking about something

If Evans is right we may find that the use/understanding distinction also has applicability to social and natural kinds, i.e. we might find a parallel distinction between:

- using a term that has a social or natural kind in its extension; and
- thinking thoughts about the kind

So far as I can tell, Crane's response to Burge's arthritis thought experiment relies on the use/understanding distinction. There are of course significant differences in dealing with proper names and social kind terms and I will discuss some of these differences in Section 2.6.

2.2 Evans's distinction between using a name and understanding a name

One type of referring term that Evans explores applying the use/understanding distinction to, is proper names. To fully appreciate Evans's position on proper names (as it is presented in *The Varieties of Reference*) requires one to see that there are two distinct theories at play:

Theory 1: a theory concerning how the referent of a name is determined – what object a person using the name would be referring to; and

Theory 2: a theory concerning what is required for a person to understand the name in such a use, i.e. what is required to entertain a thought about an object (and hence the referent of the name used)

The particular theory that Evans holds about Theory 1 is not really critical here. What is critical is the fact that he distinguishes Theory 1 from Theory 2. I will thus provide the briefest of sketches of Evans's position with respect to Theory 1. Evans proposes a broadly Kripkean (1980) causal theory. Kripke, however, did not say very much about how names operate once they have been introduced. Evans here distinguishes between 'producers', who have demonstrative encounters with the object named and 'consumers' whose use of the name is not backed up by such demonstrative encounters. Consumers join the practice on the strength of what is essentially descriptive information and rely on the producers to 'fix the reference' when they use the name. Let's call this the Kripke-Evans theory of the referent-in-use of a name (the K-E theory for short).

We can now turn to Theory 2 above. According to Evans, in order to *understand* a referring expression[9], the hearer must "link up the utterance with some information in his possession" (p. 305), that is, form an 'information-based thought', which requires the use of the subject's information system. An information-based thought is object-invoking (pp. 326-31) – i.e. of such a kind that it simply could not exist in the absence of the object (or objects) which it is about (p. 71) (what I will call a genuine singular thought[10]). According to Evans proper names are "perhaps the most reliable indicators that an information-invoking interpretation is intended", i.e. you are expected to access some information in your information system, regarding the thing named. Thus we find the following:

[9] that is not a 'descriptive name' (1982, p. 31)

[10] What Evans calls a 'Russellian thought' (p. 72)

Extract N

> …I think it will be universally acknowledged that understanding a use of a proper name requires one to go beyond the thought that the speaker is referring to some person knows as NN, and to arrive at a thought in the thinking of which one actually thinks of the object in question… (p. 398-9).

He concludes that “the single main requirement for understanding a use of a proper name is that one think of the referent” (p. 400, my italics).

Evans believes that failure to appreciate the use/understanding distinction is one of the sources of confusion over naming. Evans draws the problem out by distinguishing two distinct notions of the “intended referent of a use of a name” (p. 402):

Extract O

> one in which the intended referent is determined by determining which name-using practice a speaker manifested the intention of participating in…and one in which the intended referent is the object which the speaker is *aiming at* with his use of the name. Full understanding of a use of a name requires that the referent of the name be an object of the subject’s thought in the *second* sense. (p. 402).

Evans identifies three ‘modes of identification’ that can be used to discriminate an object, which is a necessary requirement of formulating a genuine singular thought: (i) descriptive, (ii) demonstrative and (iii) recognition-based. Once again, the details need not concern us. However, there is an associated subtlety in Evans’s account which is also important. Evans recognises that consumers will associate only descriptive material with the name (i.e. if they understand the name they will be formulating descriptive genuine singular thoughts about the thing named). However, the requirement on a speaker *using* a proper name is not that he have sufficient information to indicate which object he or she intends to be (taken to be) referring but rather to “indicate which name he intends to be (taken to be) using [i.e. which name-using practice he is participating in]”. (p. 384). The result is that Evans holds that it is theoretically possible for a consumer to successfully use a name even if all the information in his or her possession regarding the referent is false (provided the misinformation is widespread in the practice). Here is a relevant extract:

> **Extract P**
> As we saw earlier, it is consistent with being able to use a name (as a consumer) that one have wholly baseless information associated with it. If the information derives from nothing at all, then someone who interprets a use of the name by invoking the information is thereby thinking of nothing; and those who associate with a name of x only a story (widely disseminated) of the doings of y are thinking of y when they interpret uses of the name by invoking this information" (1982, p. 400)

The central point for our purposes is that according to Evans, in a situation like this, if a speaker says something like SN: "N was a great orator", then the name 'N' will still *refer* to *x*, but a consumer using the name N will be *thinking of y* when he or she interprets SN. Evans realises that he has his detractors here (p. 400) but I think that his replies are compelling.

2.3 Linguistic Content and Explanatory Psychological Content

In philosophy it is generally taken that thoughts have contents and that sentences have contents. My working assumption is that the *content* in both cases is a *representation of a state-of-affairs.* Drawing on Textor's (2012[11]) discussion I take the notion of a state of affairs to entail the following:

1. States of affairs exist independently of thinkers
2. States of affairs either obtain or do not obtain
3. States of affairs can exist without obtaining
4. States of affairs involve objects and properties/relations directly

On this conception, (2) and (3) both serve to distinguish SOAs from facts: facts just are, facts do not obtain or fail to obtain and hence a fact cannot exist without obtaining (Textor, 2012, p. 17). Textor notes that not all philosophers consider the notion to be uncontentious[12]. However, I hope that this will be taken as a reasonable assumption in the circumstances.

In summary, I take it that states-of-affairs directly involve properties and particulars, e.g. the SOA of snow's being white directly involves snow and the property of whiteness. If the

[11] 'States of Affairs', Stanford Encyclopedia of Philosophy, Mar 2012
[12] See Textor, pp. 20-34, especially Fine's response on p. 26

SOA that is represented obtains then the representation of that SOA will be a true representation.

On face value, in cases in which a speaker does not mean what he or she says (where reinterpretation would be appropriate), there will be a distinction between:

1. What the speaker said in the context; and
2. What the speaker intended to say (and presumably thought) in the context

If we restrict ourselves to differences that lie at the level of reference (as opposed to differences at the level of something like sense), a difference would imply that the SOA that the speaker said something about was not the same SOA that the speaker intended to say something about (and that the speaker had in mind, so to speak). To formulate the distinction between (1) and (2) in terms of states-of-affairs represented, we could distinguish:

3. The SOA that the speaker actually represented by virtue of uttering the words in the context; from
4. The SOA that the speaker intended to represent by virtue of uttering the words in the context (which I will treat as equivalent to the SOA that the speaker had in mind)

Or, as a convenient shorthand:

5. The SOA that is *linguistically represented* (by a speaker in a context); from
6. The SOA that is *mentally represented* (by a speaker in a context)

We can now cash out the term ‘Linguistic Content’ in the thesis title as being equivalent to 5 – i.e. Linguistic Content is the SOA that is linguistically represented (by a speaker in a context). Similarly, Explanatory Psychological Content is equivalent to 6 – i.e. to the SOA that is mentally represented (by a speaker in a context).

Earlier, when examining Burge’s position in Section 1.2 I suggested that we were faced with a trade-off between psychological sensitivity and semantic stability. On the view being put forward here, Linguistic Externalism applies to the Linguistic Content – i.e. the SOA that is linguistically represented will depend, in part, on the (linguistic) meaning and reference of the words used. In the case of social terms this will depend on the relevant (external) social facts. The Linguistic Content would thus deliver semantic stability. The central question is how the Explanatory psychological content relates to the Linguistic Content if we want the Explanatory Psychological Content to be genuinely psychologically sensitive.

We can now formulate the S→M Principle (*we mean what we say) as the following claim:*

The S→M Principle: If a person utters the words 'a is F' in a given context then the SOA that is mentally represented (the SOA that the speaker intended to represent by virtue of uttering the words in the context) is the same SOA that is linguistically represented (the SOA that the speaker actually represented by virtue of uttering the words in the context).

We can now say that if we mean what we say then the state-of-affairs that is mentally represented in a context will be the same state-of-affairs that is linguistically represented in that context.

Looking back to Evans's analysis of proper names and applying this nomenclature to the orator example that we discussed in Section 2.2 we would put things thus:

- The state-of-affairs that the speaker linguistically represents is the state-of-affairs that *x* was a great orator
- The state-of-affairs that the speaker mentally represents (and intended to linguistically represent) is the state-of-affairs that *y* was a great orator.

Some would no doubt consider the notions of linguistic representation and mental representation set out above to be contentious. Internalists about mental content, for example, would point out that externalism is presupposed here: since it is presupposed that mental representations are *representation of states-of-affairs* and SOA directly involve properties and objects. In other words, if person A has a mental representation of SOA1 and person B has a mental representation of SOA2, where SOA1 directly involves object A and SOA2 directly involves object B, then Person A and Person B will have different mental representations, even if Person A and Person B were in the same narrow state.

I accept this point which, given the theoretical framework here, comes down to the assumption that mental representations have truth conditions (on face value an approach shared with many philosophers, c/f Lau and Deutsch (2010)[13]). This is one of the reasons that I have suggested that this is essentially intended as a discussion between externalists. By accepting this point I am really just granting what many philosophers have noted, that once we admit intentionality into the picture externalism should come as no surprise. For example we have Stalnaker (1999, pp. 169-170):

> **Extract Q**
>
> In retrospect, it seems that we should not have been surprised by the conclusions of Putnam and Burge. Isn't it obvious that semantic properties, and intentional properties generally, are relational properties: properties defined in terms of relations between a speaker or agent and what he or she talks or thinks about. And isn't it obvious that relations depend, in all but degenerate cases, on more than just the intrinsic properties of one of the things related. This, it seems, is not just a consequence of some new and controversial theory of reference, but should follow from any account of representation that holds that we can talk and think, not just about our own inner states, but also about things and properties outside of ourselves"

I would also like to stress that nothing in the above presupposes a distinction *in kind* between a linguistic representation and a mental representation: distinguishing (5) from (6) is consistent with there being no such distinction. What *is* assumed is that both are representations (of states-of-affairs) and both have truth conditions (in the genuine sense). To put this slightly differently, the distinction being assumed above is a distinction at the level of *what is represented* (what SOA is represented) not *how it is represented* (i.e. mentally vs linguistically). This is why I have suggested that the linguistic representation and mental representation nomenclature is introduced as a 'convenient shorthand' for distinguishing (3) from (4).

I leave it open that by examining the reasons for distinguishing (3) from (4) and distinguishing (5) from (6) in various instances one might find evidence on which to distinguish the kind of representation that a linguistic representation is from the kind of representation that a mental representation is. However, I draw no such conclusions here and nothing that I say here will depend on it.

As I mentioned at the outset, my aim is not to argue against psychological externalism but to focus on the role that the S→M Principle plays in the type of argument that Burge

[13] However, on the assumption that a representation is an object with semantic properties (content, reference, truth-conditions, truth-value, etc.), a mental representation may be more broadly construed as a mental object with semantic properties (2010, p. 1).

presents, which in turn suggests certain *limits* on what conclusions ought to be drawn from such arguments for psychological externalism.

2.4 Guarding against equivocation over expressions like 'what is said' and 'proposition expressed'

Above we found Evans drawing attention to the possibility of a gap opening up between:

- what a speaker means to say…what thought he wishes to express…; and
- what he strictly and literally says according to the conventional meanings of the words he utters

Caution is required in interpreting what Evans means by what a person "strictly and literally says according to the conventional meaning of the words". On a straight-forward reading this suggests that the meaning gap which Evans has in mind is that between the meaning of a sentence based on something like the dictionary definition of the words that are used and what the person intended to say. However, to be interesting, the notion of 'what is said' that Evans is pointing us to must be genuinely semantic in the context. It seems tolerably clear from what Evans says elsewhere that he has in mind the possibility that the genuine semantic content (i.e. content with truth conditions) of what a person "strictly and literally says" might be distinct from the thought or belief that the speaker wishes to express.

I would like to distinguish between a definitional issue and a substantive issue that comes out with the use of expressions like 'what is said', 'what is expressed' or 'the proposition expressed'. Take the orangutan example from earlier. Let's imagine the following situation:

Jill says: 'I had an orangutan for breakfast'

Where the correct name for the type of fruit drink that she had is 'Orangisun'. Here we have two states-of-affairs in the offing:

i. the states-of-affairs of Jill's having an orangutan (a type of ape) for breakfast
ii. the states-of-affairs of Jill's having an Orangisun for breakfast

Many would describe the ‘proposition expressed’ or ‘what is said’ under the circumstances as the second one above, i.e. whatever the speaker is interpreted or reinterpreted as saying. Whether one uses these phrases to refer to (i) or (ii) above is a matter of convention. However, I will be using the term to refer to (i) – i.e. to refer to what Evans would describe as what Jill “strictly and literally says” according to the conventional meanings of the words she has used in the context (a notion of content that is just as genuinely semantic or propositional as (ii)).

One of my aims in introducing the notions of the states of affairs that are linguistically represented (a linguistic representation) and the states of affairs that are mentally represented (a mental representation) is in order to mitigate against the risk of equivocation over expressions like ‘what is said’ and ‘proposition expressed’. In the Orangutan/Orangisun case we would describe things thus:

i. the SOA that is linguistically represented is the state-of-affairs of Jill’s having an orangutan (a type of ape) for breakfast
ii. the state-of-affairs that is mentally represented is the state-of-affairs of Jill’s having an Orangisun for breakfast

So the state-of-affairs that Jill mentally represented was not the same state of affairs that Jill linguistically represented.

All that this really amounts to, so far, is increased clarity over the use of terminology. However, it is because of the risk of equivocation when using expressions like ‘what is said’ and ‘proposition expressed’ that I am suspicious of using these expressions to frame and discuss the issues under examination here. The term ‘proposition expressed’ implies success by the speaker in having communicated what was on his or her mind when the speaker ‘expresses a proposition’. Expressing oneself is the aim of language and pre-philosophically having expressed a proposition implies success. As a result it seems odd to say ‘she expressed the proposition *that she had an orangutan for breakfast* but she did not really mean to express that proposition’. There is pressure either to insist that she did not *actually* express that proposition or that if she did she must have intended to. I hope to avoid this risk by using the schema described above.

The picture that emerges from the above is that in deriving the SOA that is mentally represented we begin with the SOA that is linguistically represented and reinterpret this in light of the speaker's misunderstanding (i.e. that 'orangutan' is the name of a breakfast drink). In short form:

linguistic representation + misunderstanding ---> mental representation.

The linguistic representation is not irrelevant here: without the linguistic representation there would be no basis for deriving the mental representation, so the linguistic representation has ineliminable work to do on this view.

2.5. Crane and the meta-beliefs approach to belief attribution

2.5.1 An overview of the meta-beliefs approach

Crane (1991) proposes an alternative interpretation of Burge's arthritis example. Here is a relevant extract:

> **Extract R**
>
> For beliefs to be expressed in words, they have to go via second-order beliefs about which words are the right ones for expressing which beliefs: sentences do not, as it were, just 'squirt out' beliefs (p. 18)
>
> On this diagnosis, then, Alf has a true belief, *I have tharthritis in my thigh*, a false belief to the effect that *'I have arthritis in my thigh' is the right sentence to express this belief*, and thus makes a false statement, 'I have arthritis in my thigh'. This means that though his belief is true, he says something false when he attempts to express it. This sounds paradoxical, but it becomes clear when we distinguish, as we did when discussing Putnam, between the meanings of sentences in public languages and the contents of beliefs (p. 19)

What Crane's objection comes down to is the claim that sometimes one can only ascribe a belief by ascribing a combination of two (or more) related beliefs. Crane's suggestion as to how we should interpret the patient (Alf) in the *arthritis/tharthritis* case is as follows:

1. That he believes that he has *tharthritis* (a non-standard concept) in his thigh
2. That he believes that 'arthritis' means *tharthritis* (a second-order belief about the meaning of the word used)

According to Crane the speaker was thinking about *tharthritis* and intended to say something about *tharthritis* but actually said something about arthritis. On this view the speaker has one true belief (Belief 1) and one false belief (Belief 2).

To put this in our terms here, the state-of-affairs that Alf linguistically represents is the state-of-affairs of his having arthritis in his thigh but the state-of-affairs that Alf mentally represents is the state-of-affairs of his having *tharthritis* in his thigh. The crucial assumption is allowing that belief ascriptions will sometimes entail non-standard concepts (e.g. the concept *tharthritis*).

In motivating his position, Crane cites Extract M from Evans that draws attention to the use/understanding distinction, which we suggested flows naturally from Linguistic Externalism. Earlier in the paper Crane makes the same general point: "…we should distinguish between the conventionally assigned meaning of a word in a public language, and *the concept intended to be expressed* by the user of that word" (1991, p. 11, my emphasis).   In other words, the meta-beliefs approach relies on the use/understanding distinction.

Indeed we can see the parallels here with Evans's treatment of proper names and the meta-beliefs approach (e.g. we described Evans's response to the orator example in terms of a distinction between the state-of-affairs that the speaker linguistically represents and the state-of-affairs that the speaker mentally represents).

It is important to underline the fact that the meta-beliefs approach is just that – an *approach*, or perhaps, a strategy to belief attribution. What Evans offers us is some theoretical machinery that we can use to apply the approach in a principled way to the specific case of proper names (Theory 1 and Theory 2 in Section 2.2). One challenge for social kinds lies in finding a principled way of applying the approach, i.e. formulating theories along the lines of Theory 1 and Theory 2 that would apply to social kinds. Crane seems to have some suggestions on this score and I will return to discuss this further at the end of this part.

2.5.2 The 'standard' Burgean response

As Crane points out, the 'linguistic evidence' alone (i.e. that what Alf says is false) is insufficient to decide between the meta-beliefs account and Burge's interpretation of the arthritis case study. Furthermore, since Burge grants that it would be a mistake to take claims like 'I have a hippopotamus in my refrigerator' at face value (see the orangutan example), Burge is in just as much need of a theory that accounts for that type of circumstance as his detractors are.

Burge discusses the meta-beliefs proposal as consisting of two methods of reinterpretation that are "often invoked in tandem" (1979, p. 93):

> **Extract S**
>
> One is to attribute a notion that just captures the misconception, thus replacing contents that are apparently false on account of the misconception, by true contents. For example, the subject's belief (true or false) that that is a sofa would be replaced by, or reinterpreted as, a (true) belief that this is a *chofa*, where 'chofa' is introduced to apply not only to sofas, but also to the armchairs the subject thinks are sofas. The other method is to count the error of the subjects as purely metalinguistic. Thus the patient's apparent belief that he had arthritis in the thigh would be reinterpreted as a belief that 'arthritis' applied to something (or some disease) in his thigh. The two methods can be applied simultaneously, attempting to account for an ordinary content attribution in terms of a reinterpreted object-level content together with a metalinguistic error. (p. 93)

The meta-beliefs approach is the type of approach that combines a metalinguistic error with a conceptual error: the speaker is attributed a belief about a notion that captures the misconception (a non-standard notion, e.g. *tharthritis*) and a metalinguistic belief about the relevant word (e.g. 'arthritis'), to the effect that it represents the non-standard notion in the public language.

Burge suggests that the following 'philosophical argument' lies behind the meta-beliefs approach, what we can call the argument from deviant speaker meaning:

> **Extract T**
>
> It is insisted that we should not attribute contents involving incompletely understood notions, because *the individual must mean something different by the misunderstood word than what we non-deviant speakers mean by it* (p. 100, italics in the original)

Burge exploits the fact that understanding comes in degrees to resist this claim since he urges that if we do accept this claim then we would be forced to deny the correctness of many of our ordinary belief attributions. And of course Burge points out (in Extract H) that he only needs to show that sometimes we apply the S→M Principle in the face of misunderstandings (or partial understanding).

I agree with Burge that it is not plausible to insist that we 'not attribute contents involving incompletely understood notions'. I have already indicated that Burge is surely right that many of our notions are only partially understood. This is why the S→M Thesis grants that it is correct to apply the S→M Principle when the subject has partial understanding or a misunderstanding provided that the Conditions are met. Accordingly, we can still enquire into how the meta-beliefs approach performs in cases in which the proposed Conditions are not met.

2.5.3 The meta-beliefs approach: further objections and responses

Crane suggests that in order to apply the meta-beliefs approach one must be committed to the view that "thought can be independent of public language" (p. 11) (this also seems to be Evans's assumption as evidence by Extract M). I think that it is important not to read too much into this though: one need not be committed to the idea that one could have the complexity of thought available to humans in the absence of language. In fact, in the context of this discussion, all that this claim really amounts to is *recognition that we can have thoughts that feature non-standard concepts*.

One concern that Crane himself draws attention to is that if one generalises this procedure of attributing non-standard beliefs to individuals one could explain away almost any appearance of error in a thinker's belief by attributing a different concept to him (p. 21). Of course a residual false belief about the public meaning of the misused word would remain, just as Alf would be diagnosed as holding a false belief about the meaning of 'arthritis'. Nonetheless, clearly affording such conceptual immunity to individuals would be an unacceptable result. Crane believes that the meta-beliefs approach can cope with this provided "the right distinctions are made" (p. 21):

> **Extract U**
>
> When working out what thinkers believe, we have to take into account not only the evidence of what they say or do at a particular time, but what they *would* say or do under other circumstance. That is, we have to consider which counterfactuals are true of them.

> Taking only the evidence of one utterance or one action will, of course, radically underdetermine the correct ascription of the thinker's beliefs.
>
> Such limited evidence will also underdetermine whether a thinker makes a genuine mistake or had a non-standard concept. (p. 21)

Crane's suggestion, broadly speaking is that we would look for evidence that the individual would consistently apply the non-standard concept in a way that suggested that the concept features in a theory – perhaps it would be an 'off-beat' theory (p. 22) but provided we found such consistency in application we would be justified in attributing the non-standard concept to him or her.

The suggestion is not that we should look to a person's narrow states or into his or her brain in order to determine whether they have a particular concept. Whether they have the concept or not will be determined by their overall behaviour and overall consistency in their interactions with the world (this is what two people having a different concept 'in mind' amounts to, on this view).

It is certainly true that we normally attribute beliefs in the context of ongoing interactions with people (sometimes spanning many years) and that sometimes we discover that earlier belief attributions were incorrect when a misunderstanding emerges during the course of such interaction.

Interestingly, Crane seems to suggest that we are faced with only two alternatives in such a situation: "What settles whether T *falsely believes that a is F or truly believes that a is F** is just what counterfactuals are true of T" (p. 21). In the case of the arthritis example he says

> **Extract V**
>
> He has an off-beat theory of the various ailments he has – not a very well-informed theory, but a theory none-the-less. This is why I say he has the concept *tharthritis*…what I am urging is the importance of distinguishing this belief from the belief that he has arthritis in his thigh – a belief that neither Alf nor his counterfactual twin has (p. 22)

However, from what he says elsewhere I don't think he is really committed to our options being restricted in this way. Imagine that we are faced with a situation in which Jill says "my arthritis is spreading to my muscles" but then goes on to exhibit a complete lack of consistency and coherence with respect to her notion of arthritis, i.e. saying things like "arthritis is thankfully good for the eyes" and "arthritis of the brain is the worst form".

Granted that in this case we would not attribute to her a gerrymandered non-standard concept, *Jillarthritis.* But then again, neither should our response be that she falsely believes that *arthritis* is good for the eyes. Surely a more appropriate response in such a case should be to attribute no belief to her.

At the outset I suggested that one of the costs of qualifying the S→M Thesis by Condition 3' is that one bears the cost of not applying the S→M Principle in cases in which the Condition is not met and that as a result we give up the direct account of how thoughts and language relate to one another. The concern here is that if the SOA that is linguistically represented and the SOA that is mentally represented can come apart, we may find ourselves unable to account for our ability to use language to communicate effectively. Let's examine an example involving two laypeople discussing arthritis. One of them is Alf and the other is Ralph. Both of them, by happy coincidence, believe that you can get arthritis in the muscles. They have a brief conversation that goes as follows

Alf says:	"I think that my arthritis is spreading to my muscles"
Ralph responds:	"I thought that might be the case – you should go and consult your doctor"

If we were to adopt the meta-beliefs framework we would conclude the following:

I. Alf linguistically represents a SOA that involves arthritis
II. Alf mentally represents a SOA that involves *tharthritis*
III. Ralph linguistically represents a SOA that involves arthritis
IV. Ralph mentally represents a SOA that that involves *tharthritis*

It seems that Alf has communicated a thought about a SOA that involves *tharthritis* to Ralph and Ralph has communicated a thought about a SOA that involves *tharthritis* back again. They *understand one another* perfectly well, even though they don't properly understand the public/linguistic meaning of the word arthritis. This carries the worrying implication that the actual public meaning of *arthritis* is irrelevant to what has been communicated. However, this is not strictly true since they both got their non-standard concepts from somewhere and at the source of that non-standard concept, presumably, is the standard (socially determined) concept of *arthritis*.

It is also worth noting that although the risk of miscommunication increases if one adopts the S→M Thesis (since different people will sometimes be related to the world in different ways which might result in them associating different concepts with the same term) there are some mitigating factors. We can draw these factors out by examining a slightly different situation. In this situation Alf believes that you can get arthritis in the muscles whereas Ralph believes that there is only one kind of arthritis, rheumatoid arthritis.

To draw out the first factor we can examine Exchange 1:

Exchange 1

Alf says: “My arthritis is acting up in this cold weather, my fingers are really sore today”

Ralph responds: “You should consider moving to the Mediterranean, I’ve heard that the climate there is kind to arthritis sufferers”

This is one of those contexts in which, as Burge suggests (Extract K) we want to attribute some shared beliefs to Alf and Ralph. For example, the belief that you can get arthritis in your finger joints, that the symptoms of arthritis are worse in cold weather etc. One might imagine that we ought to analyse this situation as follows:

I. Alf linguistically represents a SOA that involves arthritis
II. Alf mentally represents a SOA that involves *tharthritis*
III. Ralph linguistically represents a SOA that involves arthritis
IV. Ralph mentally represents a SOA that involves rheumatoid arthritis

However, this would be *incorrect* since the view being examined here is the S→M Thesis and under this thesis, the meta-beliefs reinterpretation strategy is only applied when Condition 3’ is not met. Accordingly, in this context it would be correct to apply the S→M Principle and conclude that:

V. Alf linguistically represents a SOA that involves arthritis
VI. Alf mentally represents a SOA that involves arthritis
VII. Ralph linguistically represents a SOA that involves arthritis
VIII. Ralph mentally represents a SOA that involves arthritis

Alf and Ralph have not miscommunicated in this case. If this is right then we can accommodate Burge's insistence that we often count beliefs as the same even when two thinkers have only partial understandings of the concepts involved. (Once again I stress that I am not making any claim here about whether a linguistic representation and a mental representation amount to the same thing.)[14]

To draw out the second factor we can examine Exchange 2:

Exchange 2

Alf says:	"I think that my arthritis is spreading to my muscles
Ralph responds:	"That's not possible, you can only have arthritis if you have symptoms in your flexible (synovial) joints" (which would more or less be true if the claim were one about rheumatoid arthritis)

Since Condition 3' is not met we would reinterpret. Applying the meta-beliefs approach we would interpret this situation as follows:

I. Alf linguistically represents a SOA that involves arthritis
II. Alf mentally represents a SOA that involves *tharthritis*
III. Ralph linguistically represents a SOA that involves arthritis
IV. Ralph mentally represents a SOA that involves rheumatoid arthritis

One might conclude that as a result Alf and Ralph are miscommunicating. However, this is not necessarily the case for in this situation it is plausible that *Ralph would (re)interpret Alf* as having mentally represented (II) or something very similar. This is because in this case Alf's conceptual commitments are in evidence in his pronouncement. It must be granted that Ralph probably wouldn't put things this way. However, as we have pointed out, Ralph does not have the non-standard concept of *tharthritis* readily available. He might think of Alf's belief as a strange belief about arthritis (actually rheumatoid arthritis!) or as a strange belief about what arthritis is. The suggestion here is that this essentially amounts to the same – it amounts to a qualification of the term arthritis (and if he were to report Alf's beliefs to somebody else in the context in which this misunderstanding was relevant, he would similarly qualify his use of 'arthritis'). Granted the result would not be perfect understanding but it would not be undiagnosed talking at cross purposes either.

[14] See my comments on p. xx

The same holds for how Alf would interpret Ralph's response: it is plausible that Alf would interpret Ralph as having mentally represented (IV) or something very similar. Again, he wouldn't put it this way (he most likely doesn't have the concept *rheumatoid arthritis*), but he would qualify what Alf means by 'arthritis' when he *interprets* what Ralph has said. In fact he would probably at this stage start to question whether either he, or Ralph, or both of them has the 'wrong end of the stick' when it comes to arthritis and start doing some further investigating in order to clarify what they are actually talking about – i.e. asking after the public/linguistic meaning of arthritis.

We must not forget that as somebody that is ascribing beliefs (or describing the mental state of another) I am limited to using the words (and concepts) that I have at *my* disposal. I may use a word that the speaker misunderstands, even knowing that he or she misunderstands it, as the simplest and least cumbersome means of describing what they believe; if the misunderstanding is relevant to the contents then I would seek to be explicit about this to avoid misunderstandings. If not, I might not bother. As we have seen, one way of doing this is to qualify the meaning of the word implicitly. These practical limitations provide extra reason to be wary of reading *correctness* off of common practice.

2.6. Some critical disanalogies between the case of proper names and social kinds

We can distinguish three types of misunderstanding that one might want to accommodate in making belief ascriptions:

Type I: purely metalinguistic errors

Type II: metalinguistic errors combined with conceptual errors

Type III: purely conceptual errors (which I take to be misunderstandings about the way the world is *other than* factors that determine the meaning of words)

In Section 2.5.2 we noted that the meta-beliefs approach assumes that a Type II error is to be accommodated in making the relevant belief ascriptions.

Interestingly the orangutan/Orangisun type of cases that Burge suggests involve "quite radical misunderstandings" (1979, p. 90) are relatively easy to account for since they are Type I misunderstandings. In these cases the speaker has a standard concept 'in mind' but selects the wrong word to stand for the concept. The orangutan/Orangisun problem is easily remedied by attributing a belief in which the wrong word 'orangutan' is replaced by the correct word 'Orangisun'.

There are other examples of misunderstanding that are similarly easily dealt with (at least on face value). Malopropisms and 'slips of the tongue' are like this as well. Here is one from Mrs Malaprop herself from Sheridan's *The Rivals* (1775): "...promise to forget this fellow - to *illiterate* him, I say, quite from your memory." (i.e. *obliterate*; Act I Scene II Line 178). Once again Mrs Malaprop is presumed to have the notion of obliteration, she has just used the wrong linguistic token to stand for it. We can get at her meaning by simply swapping 'illiterate' and 'obliterate'. The result would be somewhat less memorable, but the humour is at her expense rather than of her making.

However, it is important to see that what we would often characterise as a misunderstanding about the meaning of a word will often only be accommodated if it is treated as a Type II rather than a Type I misunderstanding. So, although Burge clearly grants that reinterpretation is appropriate in some Type I cases (e.g. the orangutan/Orangisun case), it might strike one as surprising if the only time that reinterpretation is appropriate is when we are faced with a misunderstanding of Type I.

In our earlier discussion of Burge's arthritis example we ascertained that sometimes the meaning of the word 'arthritis' would be qualified in belief attributions. In light of the discussion in the last section it now seems plausible that what this amounts to is the attribution of a non-standard concept, where such a non-standard concept will often consist in an explicit or implicit qualification of a standard concept (the misunderstanding and qualification being peculiarly relevant in the circumstances).

If this is right then it looks as if a combination of the S→M Thesis and the meta-beliefs approach might be along the right lines; the meta-beliefs reinterpretation strategy would be invoked in these more complex cases: cases in which the speaker does not have the standard concept and the non-standard concept is relevant to the content in the context.

Earlier I suggested that one of the challenges associated with applying the meta-beliefs approach to social terms was formulating theories for social kinds that paralleled Evans's theories for proper names. Perhaps we might have hoped for something like:

Theory 3: a theory concerning how the extension of a social term is determined, i.e. what determines the extension of a social term that a person uses; and

Theory 4: a theory concerning what is required for a person to understand the social term in such a use, i.e. what is required to entertain a thought about a social kind (and hence a thought about the kind the relevant social term has in its extension)

If we accept Linguistic Externalism then social practice and social facts will feature in Theory 3 (I have assumed that this part is relatively uncontentious). If we are to apply the meta-beliefs approach in a case in which somebody misunderstands a social term, then Theory 4 will also need to cover cases in which a person is attributed a belief about a non-standard social kind (in Section 2.5.3 we found that Crane had some suggestions about what factors one might take into account here).

However, there are important differences between the situation with respect to proper names and that with respect to social terms which lead me to suspect that attempts to formulate a theory, such as Theory 4, that we could apply in a principled way to determine whether a person is able to formulate thoughts about a social kind (whether a standard or non-standard social kind) is unlikely to be successful.

One of the crucial differences between applying the use/understanding distinction to proper names and applying it to social kind terms is *the role of convention*. This is most clear if we examine referring terms that refer to discrete objects, such as people. It seems pretty clear that when we refer to a person by name we use a social convention that enables somebody *else* to think about that person (one might say that referring to something just is enabling somebody else to think of that thing). However, although the referring is done by exploiting some type of convention, thinking about a person does not seem to rely on social convention. By contrast, to think about a contract requires one to have a belief about a social construct – the notion of a contract *is* a social convention. So although thinking about a person, for example, does not require knowledge of any social convention thinking about a social kind clearly does. We can come at this point from a different angle: Evans's demand for what is required to understand a referring term, for example, is to be in a position to have a thought about the object that the term refers to. Either one does understand it or one doesn't. Intuitively this does not come in degrees – at least not if we think in terms of everyday uses of referring terms (I am sure that one could construct cases at the fringes). However, in the case of social terms it is clear that in the *ordinary* cases the level of understanding that individuals have concerning such terms does come in degrees. This suggests that it is likely to be a lot easier to formulate a sharply defined theory for what is required to understand a proper name[15] than it will be for what is required to understand a social term.

This focuses us on what I take to be the deep problem underlying the S→M Thesis: the problem is that if our conclusions so far are right then what counts as understanding a social term in a given context seems to be highly context-specific. One of the key considerations as to *whether* the S→M principle is to be applied or not in a given circumstance is whether Condition 3' is met. The problem that emerges is that whether or not Condition 3' is met seems to be somewhat ad-hoc and would need to be determined on a case-by-case basis given the context. This is why I don't think that the prospects for formulating something like Theory 4 are very good.

In any event it is worth pointing out that not all reinterpretations of a person that misunderstands a social term will involve the attribution of a belief about a non-standard social term. Consider the case of a child who's knowledge of arthritis was limited to her knowing that her granny has something called 'arthritis' in her knuckles. Let's imagine that the child gives away her lack of understanding by saying:

"Arthritis would also be bad if you got it in your tummy"

Clearly we would need to reinterpret in such a case. If we spent a bit of time over the matter, we would probably conclude that the Explanatory Psychological Content we ought to attribute to her is something like:

Belief X: The type of disease that granny has in her knuckles would be bad if you got it in your tummy [i.e. a descriptive thought]

Most often we would not waste a lot of mental energy thinking all this through - we would just correct her and explain that you can't get arthritis in your tummy and she would learn a little more about what arthritis means and how to use the term correctly. One might want to say that as speakers of a language we spend quite a lot of time learning what we are saying; this is very clear in children but carries on through adulthood.

[15] Although I am not for a moment suggesting this is easy (I touch on some problems in Part 3)

Nonetheless, the attribution of a descriptive belief in this case would seem to be correct and it may be correct in many cases in which adults have a limited grasp of a social kind – or some natural kinds for that matter (I would suggest that it might be the right way to interpret any claims I made about elm trees since I honestly couldn't pick one out). If I am right this is one of the implications of admitting the use/understanding distinction: sometimes people will use a word that belongs to one semantic category (e.g. a social term or a proper name) and thus linguistically represent a SOA about a social kind or an object, without being in a position to mentally represent the SOA that has been linguistically represented. If, like the child in the situation above, only a descriptive thought would not misrepresent the person's psychological situation, then they should be reinterpreted as mentally representing a descriptive thought in the context.

One implication of this context-dependence of mental representations is that the Explanatory Psychological Content that is attributable in the context has a lot more to do with making sense of the person in the context than whether they use a descriptive term, a referring term, a demonstrative, a natural kind term or a social kind term.

2.7. Implications for the Step 1 Argument

What then are the implications for the Reformulated Step 1 argument at the end of Section 1.2? That argument turned on P4'':

P4'' If the Conditions hold then the term 'arthritis' has the public/linguistic meaning of 'arthritis' in this oblique occurrence

We have now determined that one of the crucial Conditions is Condition 3': that Alf's misunderstandings are not relevant to his intentions and expected behaviour in the communication context.

So the Step 1 Argument is to be upheld but the following claim should be rejected:

The Strong Claim: If a generally competent speaker attempts to express a belief by uttering the words 'a is F' (where 'a' is a social term) then that person believes that a is F (i.e. that person mentally represents the state-of-affairs of a's being F)

I suspect that there are some that believe that this is the claim that commitment to social externalism requires us to accept. At the very least I hope to have shown that this is not the case.

It will be helpful to draw out what I take to be the broader implications of these conclusions by examining what Hornsby (1997) calls the principle of semantic innocence; Hornsby introduces the principle by means of the following example:

> **Extract T**
>
> Assume that (1) is correct. Then the content of a speaking of (2) is the same as a content of a belief of Bill's:
>
> (1) Bill thinks that the sky is blue
> (2) The sky is blue
>
> If a person came out with both (1) and (2), then there would be something she had done twice, namely, utter English words having the content that the sky is blue. (p. 197)

Hornsby describes the principle at play here as the Principle of Semantic Innocence: "This is the principle that the words which are used in saying what someone has said (or, more generally, words in 'that' clauses following propositional attitude terms) mean and refer to what they ordinarily mean and refer to" (p. 198). The implication of the principle that Hornsby wants to draw attention to is what I will call the *Specification Claim*

> "…specifications of what people think or want or hope or fear rely on the use of words as meaning what they do when it is, for example, stated how things are" (p. 198).

The principle of semantic innocence is clearly very closely related to the S→M principle, if the following two equivalences hold

1) "specifications of what people think or want or hope" being equivalent to "mental representations"
2) "stating how things are" being equivalent to "linguistic representations"

In what follows I will assume that these two equivalences do hold. Even so, there is still a subtle distinction to be made because the S→M principle is specified in terms of states-of-affairs represented, rather than representations per se, i.e. it is a more restricted thesis (as noted earlier, it is, if you will, a principle that applies at the level of reference rather than at the level of sense). The S→M Thesis is a theory concerning when a social term stands for a state-of-affairs directly involving that social kind (one might say, when the term has the kind in its extension). If one assumes that difference in extension implies difference in sense, then if the theory holds for SOA then in those case it would hold for a theory of sense as well (and thus to the principle of semantic innocence).

A little later, Hornsby suggests that exploiting the Principle of Semantic Innocence "is a matter of relying on the fact that we may use our words having their ordinary semantic properties in attributing beliefs to another" (p. 206).

If this is right then if Bill says:

> "Elm trees are evergreens"

And I report Bill's claim as:

> "Bill believes that elm trees are evergreens"

Then when I report Bill's claim the *term* 'elm trees' has elm trees in its extension and the resultant belief that I attribute to Bill is a belief *about Elm trees*. If one accepts the S→M Thesis (this time applied to natural kind terms) and the equivalences above, then one can say that this is pretty much right provided that we make the following qualification: I am only correct in applying the principle of semantic innocence if the Conditions hold.

The resultant position is acceptance of the specification claim *in ordinary cases* (i.e. when the Conditions hold). So, in ordinary cases, it would be correct to attribute the belief that Elm trees are evergreens to Bill. However, it would not be admitted that Bill holds this belief merely because he uttered the words "Elm trees are evergreens".

On this view, the 'ordinary semantic properties' of the words we use determine the SOA that is linguistically represented. When the Conditions hold the SOA that is linguistically represented is also the SOA that is mentally represented.

Our findings here run parallel to those set out in the last section. The Strong Claim is to be rejected:

Strong Claim: If a person attempts to express a belief by uttering the words 'a is F' then the specification claim holds true

Whilst the Weak Claim is to be upheld:

Weak Claim: If a person attempts to express a belief by uttering the words ‘a is F’ and the Conditions hold, then the specification claim holds true

2.8 Residual Concerns over the S→M Thesis

It is important to stress here that if we adopt the combination of the S→M Thesis and the meta-beliefs approach then the aim of mental content ascription is not to fully and accurately represent whatever the speaker’s ‘true’ understanding amounts to but rather to ensure that no misunderstandings that might be relevant in the context are left undiagnosed or unqualified. In other words, to claim, as the S→M Thesis does, that if the Conditions are met it is ‘correct’ to apply the S→M Principle (and to infer that a qualified mental representation is equally ‘correct’) is to claim that an Explanatory Psychological Content attribution is correct if no misunderstandings that might be relevant in the context are left undiagnosed or unqualified. We have thus endorsed Burge’s caution that the aim of mental content ascription is not to capture the speakers ‘true’ understanding (see (iii) of the Introduction).

On this view, what is required of a correct belief ascription is that it is ‘good enough’ in the context. Good enough in the sense of not misrepresenting the speaker’s intentions and expected behaviour in the communication context. What this requires is not sensitivity to the speaker’s ‘true’ understanding but to misunderstandings that might be relevant in the communication context (misunderstandings that result from the individual having a limited epistemic perspective on the world). I must admit that although I think that the analysis so far is directing us towards such conclusions, it is hard to accept them without some measure of discomfort.

These residual concerns might provide the Burgean with some justification for rejecting Condition 3’: Earlier we noted that whether or not Condition 3’ is met seems to be somewhat ad-hoc. Let’s imagine that the child from the previous example has now grown up a bit and learned quite a lot more about arthritis but believes that only ‘old’ people can get arthritis (she has over-generalised from her personal experience of people that have arthritis). Since we have already ascertained that we should not demand mastery of the concept *arthritis* from an individual in order for that individual to sometimes correctly be attributed beliefs about arthritis, at what point and in what context would we, for example, want to say that the doctor's belief *that arthritis would be a painful disease to contract* and the child's belief *that arthritis would be a painful disease to contract* are shared beliefs? In other words when ought we to determine that Condition 3' has been met (i.e. that her misunderstanding (or partial understanding) is not relevant to her intentions and expected behaviour in the communication context)? What demarcation criteria are we going to rely on in order to determine whether Condition 3' has been met?

One potential response to this concern is to do away with the ad-hocness by asserting that whenever there is only partial understanding (or a misunderstanding) Condition 3' will not be met. Indeed there seem to be good independent reasons for drawing this conclusion -– if we are dealing with a conceptual (or partly conceptual) misunderstanding then it is always *possible* that this misunderstanding will come into play in the individual's subsequent behaviour, even if it does not seem to be relevant in the immediate communication context. However, if we make this move then we are back on the wheel since this line of thinking leads one to the argument from deviant speaker meaning (refer Section 2.5.2) which we agreed with Burge should be resisted.

We can expand the example of the child above to make the 'independent reasons' a little more vivid. Let's say that she is now a teenager, and she knows that arthritis covers a range of inflammatory conditions of the joints, that she knows many of the symptoms of arthritis, etc. In addition, she has also been told by her parents that you should never leave arthritis untreated (as the joints run the risk of degrading further and faster if the condition is left untreated), and she believes them. We would surely be inclined to attribute the following belief to her:

Arthritis should never be left untreated

However, she also still believes (falsely) that only 'old' people can get arthritis. The result of this could be that she herself might become aware of all the symptoms of arthritis in herself but not go to her doctor because she believes that young people can't get arthritis (we might have to imagine that she handles her own medical affairs). Given this possibility, we might ask ourselves whether it was really correct to attribute to her the belief that you should never leave arthritis untreated?

Indeed we might find evidence in this to suggest that she did not have a full understanding of the truth conditions of the belief that had been attributed to her. Our teenager in the example above would presumably grant that her belief would be true in the following possible world: a possible world in which children with inflammation of the joints need not seek treatment but anybody with arthritis should seek treatment immediately. One might take this as evidence that she does not have a full understanding of the truth conditions of the belief that anybody with arthritis should seek treatment. Which would make perfect sense since the whole point of applying the S→M Thesis is to accommodate partial understanding and a person that has only partial understanding of a social term will presumably sometimes not have a full understanding of the truth conditions of the belief that has been attributed to him or her.

We seem to be caught between two claims that both seem right:

I. On the one hand understanding comes in degrees and it seems implausible to insist that only people with an 'expert' level of understanding of a social term (i.e. what Burge calls 'mastery' of a concept) would be able to truly hold a belief about that social kind; hence the S→M Thesis should be upheld

II. On the other hand, if we attribute beliefs in the face of partial understanding or misunderstanding then the argument from behaviour will always be able to find purchase and undermine the correctness of the belief ascription; hence the S→M Thesis should be rejected (i.e. the S→M Principle is only applicable in the absence of misunderstandings and strictly, correct belief attribution should always take account of the degree of an individual's understanding or misunderstandings)

If I am right then the problem that we are coming up against has much in common with the one that Kripke diagnosed in 'A Puzzle About Belief' (1979), for both rest on the fact that the individuals whom we are seeking to attribute beliefs to have epistemically limited perspectives on the world (which is why they have the misunderstandings in the first place).

If we assume that we come out in support of (I) it follows that we must reject (II). If we reject (II) then we come back to face the concern over the ad-hocness of when Condition 3' is met and a resultant ad-hocness in belief ascriptions. One way or another we have to face some uncomfortable results.

In Part 3 I will say something about these residual problems.

PART 3

PSYCHOLOGICAL SENSITIVITY VS SEMANTIC STABILITY: THE BROADER CONTEXT

3.1 Kripke and 'A Puzzle About Belief'

Kripke (1979) cites the case of Pierre, who is inclined to utter the following sentences:

S1 'Londres est jolie'

S2 'London is not pretty'

The reason for this is that Pierre does not associate the city he learned about when he lived in France called 'Londres' with the city that he happened to move to and which he calls 'London', having learned the English name from the people that live there (p. 392).

As Kripke notes, it seems that we want to attribute two inconsistent beliefs to Pierre

Belf1 Pierre believes that London is pretty

Belf2 Pierre believes that London is not pretty

The Pierre example is a classic example of an identity confusion: Pierre does not realise that what he takes to be two distinct places (*Londres* and London) are in fact the same place. It is because of this identity confusion that "Pierre is in no position to draw ordinary logical consequences from the conjoint set of what, when we consider him separately as a speaker of English and as a speaker of French, we would call his beliefs" (Kripke, 1979, p. 396). As Kripke points out, what Pierre lacks is not logical acumen but *information* (p. 394). Since there is also clearly no principled way to admit one of the beliefs and reject the other Kripke suggests that we cannot convict Pierre of inconsistency (p. 394). Kripke's central conclusion is that "the puzzle is a puzzle" (p. 400) and he does not propose any solution. Rather, he wonders whether the normal practices of belief attribution don't break down under the strain of cases like this. He concludes: "Hard cases make bad law" (p. 402).

If this happened in real life we would surely explain to Pierre that they are the same city and he would realise that he had experienced the same place in two very different ways which had led him to believe that he had experienced two distinct places. Putting things this way naturally directs one to considerations of something like Fregean sense. However, it is one of Kripke's aims is to show that the puzzle cannot be solved by adopting what *he* calls the 'Fregean view' (i.e. the view on which names have a descriptive sense) and thus to argue that Kripkean semantics is no worse off in this regard than the 'Fregean view'. Setting aside whether Frege really was committed to sense being descriptive[16], I believe that Kripke's concerns are well-grounded. The main reason for this is that any notion of *publicly available sense* of a proper name is going to face difficulty in accounting for cases like this. Let's assume that in order to understand the name is to grasp its public sense. In Pierre's case we face a difficulty since it seems hard to argue that Pierre does not understand the word 'London' and the word 'Londres', although he came to his understanding of each independently. He seems able to express all sorts of beliefs about both (as in S1 and S2). However, by the same token, it seems very hard to argue that 'London' and 'Londres' have different senses. So the problem stands; it is not clear how we are to explain the fact that Pierre's beliefs, Belf1 and Belf2 are inconsistent with one another. It seems that either we must convict him of inconsistency or we must insist that he does not really understand the terms 'London' and 'Londres' (i.e. insist that he has not grasped the public sense of those terms).

To the (in any event dubious) response that 'London' and 'Londres' do have distinct senses because the one is a French word and the other is an English word, Kripke points out that the same problem can arise in cases within English alone, in which it would be hard to argue that the words have different public Fregean senses (e.g. the Paderewski cases) (pp. 398-399).

The structural problem here is how to account for the beliefs of people that have *a limited epistemic perspective* on (i.e. limited knowledge of) the world. Soames (2006) draws the problem out nicely:

> **Extract W**
>
> the relationship between sentences and the propositions they express is nontransparent in an important way…There are pairs of sentences S1 and S2, and contexts C, such that in C
>
> (a) S1 expresses a proposition p1, S2 expresses a proposition p2, and speaker-hearers understand both sentences, while knowing that to accept S1 is to believe p1 and to accept S2 is to believe p2

[16] In NN Frege certainly misrepresents Russell – Russell (1905) was certainly not committed to the sense of a name being a definite description: on Russell's view names were definite descriptions which were to be understood quantificationally, which obviated the need for the notion of sense

(b) p1 bears some intimate "logical" relation to p2… e.g. p1 is identical with p2… even though

(c) speaker-hearers have no way of knowing that the relation mentioned in (b) holds between the propositions believed in virtue of accepting S1 and the proposition believed in virtue of accepting S2 (p. 242)

Interpreting this from our perspective, what emerges is that if one admits that referring terms have public meanings, whether it is a direct reference theory or an indirect reference theory (e.g. a Fregean theory), then some of those meanings will bear certain logical relations to other meanings (or rather the meanings of sentences containing the respective terms will bear logical relations to one another). For example, 'London' and '*Londres'* both refer to London. If individuals are to be granted understanding of those terms without having a full grasp of the logical relations that hold between such terms (e.g. 'London' and 'Londres' referring to the same place) then it will be possible for those people to hold inconsistent beliefs.

The problem of course is that we don't demand a full grasp of these logical relations from an individual in order to attribute beliefs which can be expressed using the referring term (just as we don't demand 'mastery' of a social term). If we examined this from Evans's perspective this is just what we would find: if understanding a referring term is being able to think about the referent, as Evans says, then presumably Pierre understands the terms 'Londres' and 'London' – he can form information-based thoughts about *Londres* and London (he is thinking about London, in some or other particular way, in each case).

If we grant then that Pierre does understand these terms then we face Kripke's puzzle – how are we to reconcile this with the fact that the beliefs are inconsistent with one another?

3.2 Applying the S→M thesis to the Pierre case

Although we have said that Pierre understands the words 'Londres' and 'London' (in that he can formulate thoughts about both of them) one might want to say that there is a broader sense in which this understanding is not full or complete (i.e. that he does not have mastery of the terms), since he does not know that the terms are co-referring. Taking this view we have a case in which we can apply the S→M Thesis – albeit to proper names rather than social terms (i.e. a case of misunderstanding in the broader sense) and assess the results.

Under the S→M Thesis the S→M Principle is only applicable when Conditions (1,2 and 3') are met. On this view, the S→M principle would be applicable in many of Pierre's interactions, e.g:

S3 'London is looking surprisingly beautiful today' [In English to his English friends]

S4 'Londres is a place I've always wanted to visit' [in French to his French friends]

However, the S→M Principle would not apply if he uttered the following:

S5 'London and Londres are different places'

S5 only makes sense if one assumes that Pierre is subject to an identity confusion, i.e. that he thinks that 'London' and 'Londres' are not co-referring terms. The identity confusion is clearly relevant to what Pierre intended to communicate (and we could imagine how this belief might affect his subsequent behaviour, e.g. he may go and ask to book a flight to Londres from London). Accordingly, we can assume that Condition 3' would not be met and the S→M Principle would not apply in this instance.

Two points are worth adding here – firstly, we could construct the same problem with less spatially distributed objects, e.g. with a person or a planet (admittedly with a little more difficulty[17]). Secondly, one could easily set the problem up such that Pierre is able to have *de re* thoughts of London when he thinks of it as London and when he thinks of it as *Londres* (e.g. he may have visited a city he knows as 'Londres' with his parents when he was younger). I will assume this in Pierre's case as it focuses us even more clearly on the issues at hand.

What sort of reinterpretation is to be recommended then? Given that we have stipulated that Pierre can entertain *de re* thoughts about both London and Londres, it seems that we want to attribute the following belief set to him:

Belief set 1 (attributable to Pierre):

i. 'Londres' is the name for that city (which I visited when I was younger)
ii. 'London' is the name for this city (which I live in now)

[17] Chalmers provides an example with the use of mirrors

iii. That city (which I visited when I was younger) and this city (which I live in now) are not the same city

The problem is that this looks more like a restatement of the problem than a solution to the problem. If (i) is a *de re* belief (of London) and (ii) is also a *de re* belief (of London), then (iii) looks inconsistent with the combination of (i) and (ii). We seem to be faced with some rather stark options:

I. Resist the intuition that Pierre is not logically inconsistent and conclude that he is logically inconsistent, he just doesn't realise it; or

II. Conclude that in fact Pierre can't actually be entertaining *metaphysically de re*[18] thoughts of London/Londres

If we want to insist on semantic stability at all costs then we have to opt for (I). If we are willing to give up some semantic stability for psychological sensitivity then we could explore something like (II). And this is just what we find bi-modal semanticists doing in response to this type of problem. I will briefly discuss one such response to Kripke's puzzle: Stalnaker's (1999) version of bi-modal semantics[19].

3.3 Stalnaker's proposal

Without getting into all the details of Stalnaker's position, Stalnaker examines Kripke's puzzle cases in 'Belief Attribution and Context' (1999) and in practical terms what his suggestion comes down to is that we (the theorists) should reinterpret the proposition expressed by Pierre in light of his mistake about the way he thinks the world is – i.e. in light of his identity confusion. For all that Pierre knows, 'Londres' and 'London' might refer to different places (he believes they do) and we should reinterpret him accordingly (c/f 1999, pp 164-165)

According to Stalnaker there is a possible way the world could be according to Pierre and we need to interpret what he says in light of this possibility. Importantly, according to Stalnaker "epistemic possibilities should be understood as a subclass of the metaphysical possibilities" (2006b p. 289). This would seem to put him back on a collision course with the

[18] Recall from Section 1.2. that an attribution is *metaphysically de re* with respect to an object *o* just in case it directly attributes a property to *o*

[19] Which he calls 'diagonalisation' (1999)

inconsistent belief set above. Soames (2006) raises just this concern about Stalnaker's position. Here is his concern in a formalised argument (based on a discussion in 2006, p, 236):

P1 Pierre knows that 'Londres' refers to that place [that he visited when he was younger]

P2 If P1 is true then Pierre has *de re* knowledge (of London) that it is referred to as 'Londres'

P3 If Pierre has *de re* knowledge (of London) that it is referred to as 'Londres' then metaphysically possible worlds which are consistent with this knowledge are limited to those worlds in which (our use of) 'Londres' refers to London

C1 Metaphysically possible worlds which are consistent with Pierre's knowledge in P1 are limited to those worlds in which (our use of) 'Londres' refers to London (by P1-P3)

By analogy one can derive the same conclusions for 'London' and hence draw the conclusion that metaphysically possible worlds which are consistent with Pierre's knowledge, are limited to those worlds in which 'Londres' refers to London and 'London' refers to London. If this is right then Pierre's belief is metaphysically impossible which presumably means that he cannot actually hold it, i.e. that he is logically inconsistent.

Stalnaker's response rests on a particular interpretation of what a metaphysically *de re* belief amounts to. According to Stalnaker, Soames's argument "rests on some controversial assumptions about de re belief…and about what it means to know what or who something or someone is" (2006b p. 290). Stalnaker's response is to accept the demonstrative knowledge attributions above (i.e. that Pierre knows that 'Londres' refers to this place and that 'London' refers to that place) but reject that it follows from this that Pierre's beliefs are metaphysically *de re* beliefs, i.e. accept P1, but reject P2. Here is a relevant extract:

> **Extract X**
>
> On my view, a de re belief attribution is correct when one can correctly and determinately describe the world according to the believer as a function of the individual [that the de re belief is about]. What this requires is not some intimate acquaintance relation, but only that there be a unique candidate. In the standard puzzle cases, there may be no fact of the

> matter about which of two distinct individuals in the world according to the believer is identical to a given individual in the actual world…. (2006b, p. 293)

According to Stalnaker, it does not follow from the demonstrative knowledge attribution in P1 that Pierre has *de re* knowledge of London. He holds that we can consistently attribute the knowledge in P1 to Pierre and include in the worlds that are metaphysically possible (from his perspective), metaphysically possible worlds in which 'Londres' refers to some city other than London.

Stalnaker's response is linked to his view that semantics is not strictly constrained by the syntactic or semantic category of a referring term:

> **Extract Y**
>
> It has been emphasized by many philosophers that referring is something done by people with terms, and not by terms themselves. That is why reference is a problem of pragmatics, and it is why the role of a singular term depends less on the syntactic or semantic category of the term itself (proper name, definite description, pronoun) than it does on the speaker, the context, and the presuppositions of the speaker in the context (1999, p. 44)

As I understand Stalnaker's response it rests on two distinct but related claims:

Claim 1: a *de re* belief attribution is correct if and only if one can correctly and determinately describe the world according to the believer as a function of the individual

Claim 2: what determines whether one can correctly and determinately describe the world according to the believer as a function of an individual, is not the semantic or syntactic category of the referring term used, nor the intimacy of the acquaintance relation between the believer and the individual, but rather the context and the presuppositions of the speaker in the context.

So, not only does Stalnaker have a particular view on the conditions required to make a *de re* belief ascription, he also has a particular view on what determines whether those conditions are met, which depends in large part on the context in which the belief is expressed and/or attributed.

On this view, it would be inappropriate to attribute a *de re* belief about London to Pierre when interpreting his claim in S5, because the requirements of Claim 1 would not be met (but in other circumstance a *de re* belief attribution might be appropriate).

If these claims are an accurate representation of Stalnaker's position then according to Stalnaker, a person's being intimately acquainted with an object, no matter how strong the form of acquaintance, is not sufficient grounds to attribute a *de re* belief about that object to that person. Conversely, neither is any particular strength of relation necessary in order to attribute a *de re* belief about an object to a person.

Soames (2006) provides an example that puts pressure on this implication. In the example, Soames has a person holding a paperweight and wondering what it is made of (the fact that the paperweight is actually made of wood being a classic case of the necessary a *posteriori*). Here is Stalnaker's discussion of this example:

Extract Z

> Here, I agree, it seems intuitively, that it is the singular proposition that this particular object is made of plastic that is compatible with my knowledge…But the composition of the paperweight is essential to it, so there is no possible world in which this specific object is made of plastic. How can what is, for me, an epistemic possibility be represented by a genuine possible world? Consider the following possible world, which I think Soames will agree is metaphysically possible: I am sitting in Soames's office, holding his plastic paperweight in my hand, wondering (just as I am in the actual world) what it is made of. This possible paperweight is a different object from his actual paperweight, though it looks and feels just like it. It does not seem unreasonable to think that a possible world of this kind is compatible with my knowledge. The two-dimensional strategy allows us to reconcile the assumption that it is with the judgement that it also is right to say that I know that it is *this* paperweight whose composition I am wondering about. (2006b, pp. 293-294)

Although this example is based on ignorance of an essential property rather than an identity, from Stalnaker's perspective the cases are analogous[20]. Given Leibniz's law we should expect this: if it is an object's essential properties that determine its identity then 'this' paperweight with different essential properties must be, as Stalnaker says above, a "different object" (ibid).

[20] i.e. which paperweight - the plastic one or the wooden one - are we to identity with the actual paperweight in the possible worlds?

Given that Stalnaker would withhold a *de re* belief attribution in this case I am not sure that it is really right to claim, as Stalnaker does, that we have effected a *reconciliation* with the claim that I know that it is *this* paperweight that I am wondering about. It seems to me that it is rather a case of us needing to change our perspective and resist the intuition that it is *strictly* this paperweight that is under discussion.

This is not to say that I consider this grounds to reject Stalnaker's position. We need to assess this in light of Kripke's (1979) analysis. One of Kripke's aims is to show that the puzzle cannot be solved by adopting what he calls the 'Fregean view' and we have upheld this conclusion. This is why he insists that "the puzzle is a puzzle" (p. 400). The clear implication here being that some sort of concession regarding our intuitions is going to be required. The concession Stalnaker asks of us is to resist the intuition that it is *strictly* this paperweight that is under discussion.

Although it may seem odd that some other object is being invoked in these cases, we must bear in mind that Pierre, for example, clearly intends to express a claim of *non-identity* (and one can say that an assumption of non-identity is implicit in the paperweight case). One obvious way of accommodating this is in terms of two (possible) objects. We could say that what the example shows us is that at least some types of modal claim simply do not make sense construed as *de re* claims.

This is why we see Stalnaker proposing a relaxing of our strict semantical/syntactical categories (recall claim 2); at least in so far as they contribute to determining what proposition is expressed in a given context. He is not saying that we should not recognise such semantic and syntactical categories; clearly interpretation of what has been said and reinterpretation begins with the strict semantics. The suggestion is that when this results in conflicts with what a person is clearly trying to express then some reinterpretation is required.

3.4 Conclusions we can draw from 'A Puzzle About Belief'

We can summarise the key points from the previous discussion as follows:

1) The puzzle cases are genuine puzzles and the notion of publicly available objective Fregean sense does not resolve these puzzles

2) Such puzzle cases are especially acute when formulated in terms of seemingly *de re* beliefs and resolving the puzzles without convicting the thinkers of inconsistency requires resisting the attribution of *de re* beliefs in these situations
3) One way of doing this is to reject the notion of *de re* beliefs altogether and insist that all such beliefs are actually essentially descriptive beliefs (and the associated possibilities merely epistemic possibilities)
4) An alternative response – the one proposed by Stalnaker, is to insist that sometimes *de re* beliefs are genuine *de re* beliefs but that in the puzzle situations the attribution of a *de re* belief is inappropriate; whether or not it is appropriate to attribute a *de re* belief to somebody depends on the context (this also rests on a particular cashing out of what a *de re* belief amounts to)

Interestingly, one of the concerns that Chalmers has raised against Stalnaker is that although he offers a demarcation criterion (Claim 1), there is a concern that it seems to be rather ad-hoc as to when a *de re* or genuine singular belief ought to be attributed . Stalnaker is unconcerned about this. As puts it “Ad hoc it may be, but I think it is not the adhocness of the theory, but of the way people understand and are able to describe the states of mind of others” (2005, weblog). Stalnaker is also concerned about the alternative, i.e. something like (3) above (2006a) and I am inclined to agree that it is hard to believe that purely descriptive information is going to carry the load in most cases.

What this examination of Pierre shows us is that even in the case of proper names and armed with Evans’s theory of what is required to understand a proper name, we were still faced with a puzzle when attributing beliefs about ‘Londres’ and ‘London’ to Pierre. The source of the problem being Pierre’s lack of knowledge due to his particular (and limited) epistemic perspective on the world. There are clearly parallels here with our question as to whether it was appropriate to attribute the belief that you should never leave arthritis untreated to the teenager, where again the source of the problem is the teenager’s lack of knowledge due to her particular (and limited) epistemic perspective on the world.

Furthermore, responses to the problem require relaxing the strict semantics, i.e. distinguishing Linguistic Content from Explanatory Psychological Content – this is in effect what the bi-modal semanticists are doing. And finally, at least one of those ways results in belief ascriptions that exhibit a degree of ad-hocness.

This gives me reason to believe (i) that the tension between the two claims set out at the end of Part 2, that I suggested both seemed right, points to a genuinely deep problem in philosophy of mind and (ii) that a response under which belief ascriptions exhibit a degree of ad-hocness ought not to be dismissed out of hand. Perhaps, to borrow from Stalnaker, the adhocness is a reflection of the way people understand and are able to describe the states of mind of others.

3.5 Semantic stability vs psychological sensitivity

I have suggested that one of the key themes and challenges here is reconciling the tension between semantic stability and psychological sensitivity. In closing I would like to touch on what I take to be responses that various philosophers have made to this problem, starting with Frege in the hope of further drawing out the significance of this problem for philosophy of mind and language.

Frege introduced the notion of sense in order to deal with the problem of informative identity statements (a problem quite closely related to the puzzle cases). Specifically, given that Hesperus and Phosphorous refer to the same object (Venus), Frege wanted to explain why the following could sometimes be informative:

Prop Hesperus and Phosphorous are identical

Frege's answer was that 'Hesperus' and 'Phosphorous' have (or perhaps are associated with) different senses or "modes of presentation" (On Sense and Reference, 1892, p. 24) of Venus. Frege thus proposes an indirect theory of reference where the word 'Hesperus' stands for a sense (a self-subsistent publicly available abstract object) and the sense determines the reference. Anybody that understands the name 'Hesperus' grasps the publicly available self-subsistent abstract object that is the sense of Hesperus.

Many have raised concerns about this proposal. For example, both Evans (1982, p. 22-25) and Dummett (1986) have argued forcibly against the notion of sense as a self-subsistent abstract object. However, I would rather come at the problem slightly differently: by focusing on the trade-off between cognitive sensitivity and semantic stability. Frege introduced Sinn in order to capture an expression's 'cognitive value' (we introduced Frege's test for difference of *Sinn,* the Intuitive Criterion of Difference or ICOD, in Section1.2)

However, as Burge points out, Frege also wants Sinn to play a semantic role. Burge (1979) distinguishes these two distinct roles for Frege's notion of Sinn as follows (p 429):

i. An epistemic/psychological role, i.e. accounting for possible differences of belief (as distinguished by the ICOD); and

ii. A semantic/linguistic role, i.e. to uniquely determine the referents of linguistic expressions (and mental or linguistic acts)

The problem for Frege is that it is not clear that Sinn can be semantically stable (i.e. play role (ii)) and be sufficiently psychologically sensitive to meet the ICOD. If we accept Frege's assertion that it is an objective property of expressions of a language that they have a definite sense then if two individuals are competent speakers of their language we can be assured that when A uses language to communicate a Thought to Person S, person S will grasp the same Thought that A had in mind. As Evans points out, based on what Frege says about Sinn, what this assurance amounts to is that when A and S communicate successfully with one another "A will not have a thought distinct, by the Intuitive Criterion of Difference, from S's" (p. 22, my emphasis). However, according to Evans, when one examines uses of proper names in natural language, it "seems impossible to force them into this mould" (p. 40). He argues that it seems possible to give two individuals different but adequate introductions to a name which would result in them both being competent with the name but applying the Intuitive Criterion of Difference differently to sentences in which the name features as a referring term (p. 40). Evans's suggestion is that sense is better cashed out as *a* way of thinking about an object (rather than "in the same way", (1982, p. 316)). Once this concession is made then the notion of a single shareable publicly available sense for a term like 'Hesperus' begins to slip away.

I think that Evans is certainly right here and indeed Frege was aware of the possibility of situations like this:

> Now it is possible that Herbert Garner takes the sense of the sentence 'Dr Lauben has been wounded' to be true while, misled by false information, taking the sense of the sentence 'Gustav Lauben has been wounded' to be false. Under the assumptions given these thoughts are therefore different. (The Thought, p. 298)

The crucial point is one that Burge rightly identifies; in order for sense to be cognitively sensitive in the way that Frege wants (i.e. to play the relevant epistemic role), senses must be distinguished finely enough to match the person's particular epistemic viewpoint in a context (p. 428). Frege sought to reconcile this with his commitment to a single objective shared Sinn by insisting that we would have no such problem in a perfect language : "So long as the reference remains the same, such variations of sense may be tolerated, although they are to be avoided in the theoretical structure of a demonstrative science and ought not to occur in a perfect language" ('On Sense and Reference', p. 58).

However, it is actual natural language that we are dealing with. It seems that it is not plausible to demand that whatever is referred to be presented in a single way that is associated with every proper name. In other words, not only does Fregean Sinn not solve the London/Londres puzzle it begins to look as if *Fregean Sinn* is not available to us as a means of solving the Hesperus/Phosphorous problem of informative identity statements either.

Putnam (1975) attacked the 'traditional' (what he took to be Fregean) approach to meaning precisely on account of such theories seeking to be both semantically stable and psychological sensitive. Here is part of Putnam's introduction:

> writers on the theory of meaning have purported to discover an ambiguity in the ordinary concept of meaning, and have introduced a pair of terms – *extension* and *intension* or *Sinn* and *Bedeutung*, or whatever - to disambiguate the notion (p. 132).

Putnam notes that "[n]one of these philosophers doubted that understanding a word (knowing its intension) was just a matter of being in a certain psychological state" (p. 134). The problem with this he suggests is that the 'traditional' theory or notion of meaning that results rests on two assumptions that "are not jointly satisfied by any notion, let alone any notion of meaning" (pp. 135-136):

1. That knowing the meaning of a term is just a matter of being in a 'narrow' psychological state
2. That the meaning (in the sense of 'intension') of a term determines its extension

I do not intend to critically examine Putnam's argumentation as such – after all commitment to Linguistic Externalism has here been assumed (although not, I should note, on the strength of Putnam's argumentation). What I do wish to focus on is how Putnam trades off psychological sensitivity with semantic stability. His conclusion that ''meanings' just ain't in the head!' (1975, p.227) comes hand-in-hand with his conclusion that *meanings* *are not intensions*: "…to say, as we have chosen to do, that difference in extension is *ipso facto* a difference in meaning for natural-kind words, thereby giving up the doctrine that meanings are concepts, or, indeed, mental entities of any kind." (p. 152). Since Putnam treats the terms 'concept' and 'intension' as equivalent[21]: what Putnam seems to be suggesting is that we should de-link the notions of concept/intension from the notion of meaning: "(This shows that the identification of meaning "in the sense of intension" with concept cannot be correct, by the way)" (p. 144)

It seems that according to Putnam the only way that we can make sense of the notion of meaning is as an a-psychological notion – as what one might call 'linguistic meaning', though of course he would just call it 'meaning' since on his view we are clarifying what the notion of meaning is. Along with other two-factor theorists, he seeks to accommodate psychological sensitivity by proposing a theory where the narrow factors account for the behaviour and the broad factors determine the truth conditions. Putnam's 'two factor' theory consists of:

1. A stereotype: broadly speaking a set of descriptions that a speaker requires in order to be competent with a word and thus for token uses of the word to be assigned 'the standard extension'
2. Extension conditions: something that the world external to the subject adds to the stereotype to make it (or at least the proposition in which it features) genuinely semantic

So strictly what we find with Putnam is that his psychological states 'proper' are narrow and as a result psychological states 'proper' are not genuinely semantic. His position thus amounts to the "higher cost" position that is associated with rejecting the S→M Thesis (see (iii) of Introduction), specifically rejection of the idea that content which is psychological 'proper' is genuinely semantic. This is the result of his de-linking the notions of concept/intension from the notion of meaning. We find the following divide in Putnam (1975):

- narrow/psychological content: intensions/concepts/stereotypes (without suggesting that these are all equivalent)

[21] there must be another sense of "meaning" in which the meaning of a term is not its extension but something else, say the "concept" associated with the term. Let us call this "something else" the intension of the term (p. 134)

- broad/semantic content: meaning/understanding/knowing (again without suggesting that these are all equivalent)

It is because of this division that Putnam is unable to equate acquiring a stereotype with acquiring understanding[22] (on his view, the acquisition of a word/stereotype is just a matter of being in a narrow psychological state but knowing the meaning of/understanding a term is not just a matter of being in a narrow psychological state). Not only are psychological states 'proper' not genuinely semantic, they are not sufficient for understanding either.

Interestingly, McGinn (1989) has suggested that there is a very natural and quick extension from Linguistic Externalism for NK terms (as Putnam upheld) to psychological externalism: if we accept the principle that the concept expressed by a term is given by what it means, where the concept expressed is the content of the propositional attitude reported, then externalism must hold for the propositional attitude so reported. This move amount to re-connecting concepts/intensions with meaning. According to Crane, McGinn's argument relies on the assumption that the contents of intentional states have truth conditions (1991, p. 4), which amounts to the same. In short, Linguistic Externalism is combined with Psychological Externalism and the psychologically internalist two-factor approach that Putnam proposed is rejected. This conclusion is here endorsed since, as already noted in Section 2.3 it has been assumed here that mental representations have truth conditions.

We could reformulate McGinn's argument in terms of the S→M principle as follows:

P1 Linguistic externalism holds for NK terms

P2 The S→M principle holds for NK terms (we mean what we say)

C1 Psychological externalism holds for thoughts expressed using NK terms

Which is fundamentally the same form of argument that Burge proposed and that we have examined in some detail. We might anticipate that the same conclusions would follow, i.e. that the S→M principle would only apply when the Conditions hold.

[22] Our reason for introducing this way of speaking is that the question "does he know the meaning of the word 'tiger'?" is biased in favour of the theory that acquiring a word is coming to possess a thing called its "meaning". Identify this thing with a concept and we are back at the theory that a sufficient condition for acquiring a word is associating it with the right concept (or, more generally, being in the right psychological state with respect to it) – the very theory we have spent all this time refuting. So, henceforth, we will "acquire" words, rather than "learn their meaning" (1975, p. 167)

3.6 Two explanatory projects

Although I have not presented an argument for the claim that mental representations have truth conditions, I should say that I believe that there are good reasons to doubt that any genuinely narrow content is going to be the psychological content that features in our psychological explanations of one another's behaviour. One view I find appealing is the view that "scientific" psychology (to the extent that it is a reductive physicalist discipline) is not sufficient to explain behaviour; it is sufficient to explain bodily movements. What explains behaviour is intentional psychology. It follows that on this view "scientific" psychology and intentional psychology are different; my suggestion is that they are engaged in different explanatory projects.

I find support for this view in McCulloch (1989) who suggests, for example that "From the fact that mental states, *considered as internal mechanisms of the human individual*, should be susceptible to internalist classification, it of course does not follow that mental states, *considered as Intentional states of the human agent,* are either classified, or even classifiable internalistically" (pp. 224/225). Evans (1982) also comes close to being explicit about a proposal of this sort (p. 204)[23]. Obviously not all of these philosophers are saying exactly the same thing. However, the fact that they are all saying something similar suggests that this is a perspective that may be worthwhile exploring.

On this view, one explanatory project (what I will call the reductive scientific project) is at the level of explanation that reductive physicalists aim for. Such a level of explanation would presumably be sufficient to account for bodily movements, although such explanations would be unlikely to be expressible using ordinary language. Another explanatory project – the one that intentional psychology is generally engaged in, is to explain how people interact with the world and with others (this explanatory project extending, of course, to our explanations of ourselves). This is a level of explanation which makes essential use of things in the world. Accordingly, the types of mental states that are

[23] Evans (1982) suggests that although a person and his or her Doppelganger would share a "more general disposition" (p. 204) because they have the same brain states, he queries what arguments there are for "holding that mental state must be identified with, or individuated in terms of, dispositional states of the general sort [of the sort that result from having the same brain state] rather than dispositions of the more specific sort" (p. 204).

invoked in such explanations are irreducible (irreducible if the reduction is intended to be in terms of a person's body and brain). This, it seems to me is the perspective from which we assess the thought experiments that are discussed in this thesis.

I am not suggesting that psychology as a scientific discipline does not make use of the notions of intentional psychology (for practical purposes I think it must if the scientists are to get any purchase on their subject matter). The proposal then is that in any given case there will be two different types of psychological explanation in the offing; a narrow psychological explanation and a broad psychological explanation (and I would urge that it is the latter type of explanation that is our principal concern as philosophers of mind and language).

3.7 Linguistic Externalism again

I have presupposed that Linguistic Externalism holds. It might be objected that surely there is no meaning of any sort in the absence of people – in particular, in the absence of people expressing thoughts and beliefs. By starting with the notion of public/linguistic meaning that is independent of individual speakers aren't we just starting in the wrong place? As I suggested earlier I recognise that there might be other ways of going about examining the relationship between thought and language. I have begun with Linguistic Externalism because that is where Burge's argument begins. Nonetheless, digressing very briefly I will add two points here. On the one hand, I do think that most people have a pre-theoretical notion that our words and sentences mean something, independent of what we (as individuals) think they mean or intended to communicate using those words and sentences. On the other hand, truth and judgement are intimately connected and the very idea of a sentence having truth conditions rests on the idea that a *person* would judge those contents to be true or false. It thus seems odd to say that sometimes the SOA that is linguistically represented will *not be mentally represented by anybody* in the context (no thought will concern the SOA linguistically represented). To my mind what we need to grant here is that if one adopts this position then the notion of linguistic/public meaning is a highly idealised notion. The notion of the state-of-affairs linguistically represented is something like the state-of-affairs that an omniscient English speaker would mentally represent (or perhaps assay) if he or she heard the utterance in the context but did not take account of any idiosyncratic beliefs that the speaker might have had. It is important here to distinguish, as Evans does (see Extract O), between a person's signalling an intention to use a particular linguistic counter (i.e. word) versus what that person intended to communicate using that linguistic counter. If one is to determine linguistic meaning in the context one must, of course, take the first intention into account. It seems to me that it is quite plausible that when we express thoughts we seek to use the right linguistic counters and that we do so because we have the notion of linguistic meaning as a target.

Repeating an observation I made in respect of the Burge discussion, it is crucial to appreciate that speakers attempt to use their words, and generally believe that they are using their words, in accordance with the public meaning of those words (i.e. that their idiolects are aligned with the public language with respect to particular utterances). The reason that we can communicate at all is because we share a public language. It of course does not follow that any of us fully understand the language we share. What does follow is that because we use language to communicate we strive to align our idiolects with the public language. When we discover that we misunderstood the meaning of a word we seek to amend our use of the relevant word in the future.

3.8 Conclusions

A key theme that has emerged here is that of reconciling the tension between semantic stability and psychological sensitivity. I have suggested that this is a problem with a good pedigree which can be traced at least back to Frege, and that Frege's attempt to solve the problem through introducing the notion of objective publicly available *Sinn* did not resolve the problem. Putnam's attack on traditional semantics correctly identifies the problem. However he, like other two-factor theorists, seems to have concluded that the problem is irreconcilable and that psychological content 'proper', that explains behaviour, is not genuinely semantic. Like McGinn and various others we have assumed here that psychological content proper is genuinely semantic. However, that merely brings the problem back to centre-stage again. Burge's suggestion is that we should de-emphasise the psychological sensitivity of belief ascriptions. However, the arguments from behaviour and close examination of his thought experiments and the Conditions suggest that some rather heavy qualifications are required here. The suggestion here, at bottom, is that there is a middle road that should be seriously explored – the combination of the S→M Thesis and the meta-beliefs approach. The price is that our belief ascriptions exhibit a degree of ad-hocness. Arguably that is not a weakness with the approach so much as a reality of belief ascriptions in a world in which subjects have limited epistemic perspectives.

My conclusion is that the S→M Thesis (in combination with the meta-beliefs account) supports further work and refinement.

REFERENCES

- Burge, T. 1979. 'Individualism and the Mental'. In P. French and T. Uehling, eds. *Midwest Studies in Philosophy Volume IV – Studies in Metaphysics.* Minneapolis: University of Minnesota Press.

- Burge, T. 2005. *Truth Thought and Reason, Essays on Frege,* Oxford University Press, Oxford.

- Crane, T. 1991. 'All the Difference in the World'. *The Philosophical Quarterly,* 41: 1-25.

- Dummett, M. 1986. 'Frege's Myth of the Third Realm' in *Frege and Other Philosophers,* Clarendon Press, Oxford.

- Evans, G. 1982. *The Varieties of Reference*, Oxford, Clarendon Press.

- Frege, G. 1892. 'On Sense and Reference' As reprinted in A.W. Moore (ed.) Meaning and Reference. Oxford: Oxford University Press.

- Frege, G. 1918-19. 'The Thought: A Logical Enquiry', *Mind* vol 65, 259: 289-311.

- Fodor, J. 1980. 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology'. Reprinted in D. Rosenthal, ed. *The Nature of Mind*. Oxford: Oxford University Press

- Hornsby, J. 1997. *Simple Mindedness: In Defense of Naïve Naturalism in the Philosophy of Mind*. Cambridge, Massachusetts: Harvard University Press.

- Kripke, S. 1979. 'A Puzzle about Belief' in *Meaning and Use, A. Margalit (ed).*

- Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

- Lau, J and Deutsch, M, 2010. 'Externalism About Mental Content', The Stanford Encyclopedia of Philosophy (Fall 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2010/entries/content-externalism/>

- Loar, B. 1988. 'Social Content and Psychological Content'. Reprinted in D. Rosenthal, ed. *The Nature of Mind.* Oxford: Oxford University Press.

- McGinn, C. 1989. *Mental Content*. Oxford: Basil Blackwell.
- Patterson, S. 1990. 'The Explanatory Role of Belief Ascriptions'. *Philosophical Studies, 59(3): 313-332*
- Putnam, H. 1975. 'The Meaning of Meaning'. Reprinted in *Mind, Language and Reality*. London: Cambridge University Press.
- Russell, B. 1905. 'On Denoting' Mind, New Series, Vol. 14, No. 56, pp. 479-493.
- Soames, S, 2006. 'Understanding assertion' in Thomson, J and Byrne, A (Eds). *Content and modality: Themes from the Philosophy of Robert Stalnaker,* Oxford University Press, Oxford
- Stalnaker, R. 1999. *Context and Content*, Oxford University Press, New York
- Stalnaker, R. 2006a. 'Assertion Revisited: On the Interpretations of Two-Dimensional Modal Semantics'. In Garc (ed.), Two-Dimensional Semantics (pp. 293 – 309). Clarendon Press, Oxford
- *Stalnaker, R. 2006b. Responses* in Thomson, J and Byrne, A (Eds). *Content and modality: Themes from the Philosophy of Robert Stalnaker,* Oxford University Press, Oxford.
- Textor, M. 2012. 'States of Affairs', The Stanford Encyclopedia of Philosophy *(Summer 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/states-of-affairs/>.