Semantic Internalism and Externalism
in the *Oxford Handbook of the Philosophy of Language*, ed. by Barry C. Smith and Ernest Lepore. Oxford University Press 2006. pp. 323-40.

Katalin Farkas
Central European University, Budapest

1. Three claims about meaning

In a sense, the meaning of our words obviously depends on circumstances outside us. 'Elm' in English is used to talk about elms, and though I could decide – perhaps as a kind of code – to use the word 'elm' to talk about beeches, my decision would hardly change the English language. The meaning of 'elm' depends on conventions of the language speaking community, and these are certainly beyond my control. In this sense, no-one will disagree that meaning is determined by factors outside the individual. At the same time, it seems that it is up to me what *I mean* by my words; and in fact, the meaning of a word in the language is simply a result of what most of us mean by it. Another way of putting this point is that even if the meaning of an expression is determined by social agreement, grasping the meaning of the word is an individual psychological act. I may grasp the usual public meaning correctly, or I may – willingly or accidentally – mean something different by the word, but it looks that meaning in this sense depends entirely on me.

It is also plausible to assume that in some sense, our physical environment contributes to what our words mean. If I am right in assuming that before Europeans arrived to Australia, English had had no word which meant the same as the word 'kangaroo' does nowadays, this is easily explained by the fact that people at that time hadn't encountered kangaroos. However, a further question is whether it would have been *possible* to have a word with same meaning, if kangaroos had never existed, or no-one had ever met them. And it seems the answer is yes. You can learn what 'kangaroo' means without ever having seen a kangaroo, say from descriptions or drawings; and descriptions and drawings can be made about non-existent creatures. If this were not so, we couldn't have words like 'yeti' or 'unicorn'. Thus the existence of kangaroos, though in actual fact did play a role in a word acquiring its meaning, is not necessary for having a word with this meaning. This brings us to our first claim: meaning is independent from – social and physical – factors outside us.

The function of certain expressions in the language is to refer to things, and expressions refer to things in virtue of their meaning. This is so obvious that it almost defies explanation or supporting argument. If we use the word 'Morning Star' to talk about the Morning Star, what could possibly determine the fact that the expression refers to the Morning Star – rather than say, to the Mont Blanc –, if not its meaning? What we learn when we learn the meaning of the expression is precisely that it is used to talk about

a certain thing. And if two expressions like the 'Morning Star' and the 'Mont Blanc' refer to different things, this must be in virtue of the difference in their meanings.[1] The claim applies even to direct reference theories, that is, to theories which regard a name's reference its only semantic feature – here the only relevant semantic feature of a name determines its reference by being identical to it. So our second claim about meaning is: meaning determines reference.[2]

The truth-value of a sentence is determined by what the words in the sentence mean and how things are in the world. As Quine says in "The two dogmas of empiricism", the sentence 'Brutus killed Caesar' is true, but it would be false if either 'killed' meant the same as now 'begat' means, or if the world had been different in certain ways. It would also be false, we may add, if for example the name 'Caesar' referred to Octavius, instead of Caesar. We can see how the meaning of the words – by determining their references, by expressing certain relations or activities – collaborate to determine the conditions under which the sentence is true; and if those conditions obtain in a world, the sentence is true. We shall call the aspect of the meaning of a declarative sentence which is responsible for its truth-conditions its 'content', and our third important claim about meaning – which parallels the second – is that the content of a sentence determines its truth-conditions.[3]

2. The Twin Earth arguments

We have introduced three plausible claims: that meaning depends on the individual; that meaning determines reference; that the content of a sentence determines its truth-conditions. However, in an influential paper published in 1975, Hilary Putnam argued that the first statement is incompatible with the second two.

Putnam's particular case is very well known by now, but let us state it for the record. We are asked to imagine a planet we may call 'Twin Earth', which is just like Earth in most respect, with one difference. The transparent, colourless, odourless liquid which flows in the rivers of Twin Earth, and which people on Twin Earth who speak a language which sounds just like English call 'water', is not H_2O, but has a different complex chemical composition, which we shall abbreviate as XYZ. H_2O and XYZ are distinguishable only by using sophisticated chemical analysis, but in normal circumstances they look, smell and taste the same. Putnam's first contention is that XYZ is not water. If a spaceship travelled from Earth to Twin Earth, travellers from Earth may

[1] McCulloch 1995: 66 and McDowell 1992: 309 give similar expressions to the obviousness of the idea that meaning determines reference.

[2] Of course, there are names like 'Pegasus' which do not refer to anything, but this is also a consequence of their meaning; compare 'Pegasus' and 'Bucephalus'.

[3] This formulation allows for the identification of the meaning of a declarative sentence with its content, and that, in turn, with its truth-conditions. In this case, the determination is trivial. However, it also allows the meaning to be richer or more fine grained than truth-conditions or whatever determines truth-conditions; and it also allows regarding truth-conditions as states of affairs, and distinguish them from features of abstract entities or mental states.

think first that Twin Earth has water; later, when chemical analysis is done, they would find that they had been wrong. Since XYZ is not water, our word 'water' does not refer to XYZ, and parallel considerations would show that the Twin Earth word 'water' does not refer to H_2O. [4]

Next we are asked to go back in time to say 1750, when the chemical composition of water was not known. Putnam maintains that the word 'water' had the same reference back then as it has now; the subsequent discovery that water is H_2O hasn't changed the meaning and hence the reference of 'water', but simply taught us something about the stuff we have been calling 'water' all along. If this is right, then already back in 1750 the word 'water' as used on Earth referred only to H_2O, and not to XYZ. And similar considerations about Twin Earth would show that their word 'water' referred only to XYZ, and not to H_2O.

Now enter Oscar, an inhabitant of Earth who lived in 1750, and suppose that by some cosmic coincidence, there lived someone on Twin Earth, call him 'Twin-Oscar' (known to his friends simply as 'Oscar'), who was an exact, atom-by-atom replica of Oscar, and shared the same history throughout his lifetime. Oscar and Twin Oscar are internally the same.Two remarks should be made here. First, we set up the Twin Earth scenario in the usual way, assuming that Oscar and Twin Oscar are internal physical duplicates. Internal *physical* sameness entails internal sameness only if Oscar and Twin Oscar are entirely physical entities, an assumption many philosophers are happy to accept. However, if someone thinks that Oscar and Twin Oscar have also non-physical properties, the thought-experiment has to be modified by offering a different notion of internal sameness. Another problem is that Oscar's body contains a significant amount of H_20, and if there is no H_2O on Twin Earth, Twin Oscar cannot be a physical duplicate of Oscar. The usual answer to this is that we could easily choose another substance which is not to be found in the human body.[5]

Oscar refers exclusively to H_2O by 'water', and Twin Oscar refers exclusively to XYZ by 'water'. If we retain the assumption that meaning determines reference – that is, sameness of meaning implies sameness of reference, and consequently difference in reference implies difference in meaning –, then the meaning of 'water' is different for Oscar and Twin Oscar. This means, however, that internal sameness does not imply

[4] A scenario similar to Twin Earth is invoked by Strawson (in Strawson 1959: 20): another sector of the universe reproduces this one. Strawson's point is that since these sectors agree descriptively, mere description is not sufficient to secure particular reference. Evans 1982: 45ff includes further discussion of this idea.

[5] A suggestion for a notion of internal sameness which is applicable to non-physicalist theories, and at the same time deals with the problem of the human body's containing H_2O, is found in Farkas 2003. The idea is, briefly, that we could have a perfectly good Twin Earth argument for example about a disease, which is found only in the brain. Therefore the boundary between the internal and the external should not be drawn around the brain, or the body, or the skin, but should be formulated in terms of the subject's perspective. This has further consequences to the issue of self-knowledge.

sameness of meaning; meaning depends on factors outside the individuals.[6] Hence Putnam's famous slogan: 'Meanings ain't in the head.'

Let us run a slightly different version of the argument. Oscar and Twin Oscar are internally the same. When Oscar says 'Water quenches thirst', his sentence is true if and only if H_2O quenches thirst. The same sentence uttered by Twin Oscar is true iff XYZ quenches thirst. Thus the truth-conditions of their sentences are different. If we retain the assumption that content determines truth-conditions – that is, sameness of content implies sameness of truth-conditions, and consequently, difference in truth-conditions implies difference in content –, then the content of the sentence 'Water quenches thirst' is different for Oscar and Twin Oscar. This means, however, that internal sameness does not imply sameness of content: the content of (some of) our sentences depends on factors outside the individual. And this is the view known as semantic externalism.

In the version of the Twin Earth argument just presented, we saw that – contrary to our initially plausible statement about meanings – features of our physical environment may play a constitutive role in determining the meaning of our words. Putnam offers an another argument to show that the other external feature we discarded originally, the social community, can have a similar role. To use Putnam's example: suppose that Oscar knows that elms and beeches are some sort of deciduous trees, but he has no further knowledge of the subject. Some people in Oscar's linguistic community – the 'experts' – know what the difference between an elm and a beech tree is, but Oscar is not one of them; he simply uses these words with the assumption that *someone* must know what the difference is. This phenomenon is called 'the division of linguistic labour'. It is still plausible, however, that the word 'elm' in Oscar's idiolect refers to elms only, and not to beeches; if he said 'there is an elm tree in my back garden', he would speak truly if and only if there was an elm tree in his back garden. Now imagine that Twin Oscar's linguistic community has the word 'elm' and 'beech' too, but as it happens, they are swapped: on Twin Earth, 'beech' refers to elms, and 'elm' refers to beeches. Thus we find Oscar and Twin Oscar in the familiar situation: despite their internal sameness, their words 'elm' refer to different things; the truth-conditions of their sentences 'there is an elm tree in my back garden' are different.[7]

---

[6] It may be objected that difference in meaning for internally qualitatively identical subjects does not entail that the difference is due to some *outside* factor: it could be due to the mere fact that we have two different individuals; for example, suppose that 'I' means something different for everyone simply because we are different individuals. (There could be an analogous view about intrinsic properties: 'being identical to Oscar' and 'being identical to Twin Oscar' could be regarded as an intrinsic property, which internal duplicates do not share.) For a defense of the view, see Searle 1983. chapter 8; for discussion and criticism, Newman, *forthcoming*.

[7] For similar arguments about belief contents, see Tyler Burge's classic paper, Burge 1979. Although many defenders of externalism see the argument from natural kind terms and from the division of linguistic labour as making similar points, the two arguments are independent. For a view which favours the first, but not the second argument, see McCulloch 1995, esp. pp. 175-181.

Sometimes the moral of Putnam's arguments is expressed as the *denial* of the doctrine that meaning (or sense) determines reference – where meaning is conceived as determined by internal facts. But this is not a good way of putting the matter. That *reference* is external or 'outside the head' is hardly a surprising claim; the view we are considering is interesting because it states the externality of *meanings* or *contents*. Twin Earth arguments proceed by first pointing out that references are different for internally identical subjects, and then arguing further that difference in reference implies a difference in meaning. And this implication holds only if there is a determinate reference belonging to a meaning. So in fact, the assumption that meaning determines reference (or the parallel assumption that content determines truth-conditions) is crucial to these type of arguments for externalism.

Although Putnam's original argument was about meanings, a further important step in the history of the debate was when the externalist thesis was extended to mental contents.[8] Mental states like beliefs are similar to declarative sentences in that they also have semantic features: they can be true or false, and thus have truth-conditions, and can be about certain things in the world. The characteristic of a belief which is responsible for its semantic features is called its content, and just like in the case of sentences, content determines truth-conditions. Moreover, it is natural to assume that the content of a sentence is the same as, or at least stands in a one-one correspondence to, the content of the belief we express when using that sentence. Using this link between thought and language, it is easy to extend the moral of the Twin-Earth argument to mental content. Since Oscar's belief which he expresses by saying "Water quenches thirst" is true iff H_2O quenches thirst, and Twin Oscar's belief which he expresses by using the same words is true iff XYZ quenches thirst, the truth-conditions, and hence the content of their respective beliefs, and hence the beliefs themselves, are different.

As the simple application of the same argument for an externalist conclusion about meanings and mental contents shows, the issues raised by these two varieties of content externalism are largely the same. Externalism is principally a view about the conditions for truth and reference, and invokes the same considerations whether it is the truth of a sentence, or the truth of a belief is in question.[9] Notice also that our initial formulation of the problem about meaning has already involved a reference to mental states: Putnam characterised the internalist position as the claim that grasping the meaning of a word is an individual psychological act. However, in what follows, we will keep language in focus, and merely indicate connections with questions about the mind.

3. Reference – same or different?

[8] See for example Burge 1982. On various options of how the argument may be extended, see McDowell 1992.

[9] Though externalism about other aspects of mental states – for example, externalism about attitudes, as it is developed in Williamson 2000 – has no parallel in the linguistic case.

The argument presented so far has centred upon the following claims:

- Oscar and Twin Oscar refer to different things by the term 'water'; the truth-conditions for their sentences 'Water quenches thirst' are different.
- Meaning determines reference, content determines truth-conditions

Let us now consider these steps and possible objections in more detail.

The first statement, even if it is not immediately intuitively obvious, is supported by the theory of natural kind term advocated by Kripke and Putnam.[10] This theory can, of course, be criticised and ultimately rejected, and it can be claimed that 'water' has the same reference on Earth and Twin Earth.[11] But this in itself can be regarded as a conclusive refutation of externalism only if there aren't any *other* types of expressions which exhibit similar phenomena. In fact, there are such expressions. Suppose that long before Twin Oscar's time, there lived a philosopher on Twin Earth, called 'Aristotle', whose life and influence exactly paralleled those of our Aristotle. When Oscar and Twin Oscar use the name 'Aristotle', they clearly refer to different individuals; to say that both of them refer to both philosophers – and to any other philosopher on other planets whose life was similar – is very implausible. (If we accept Kripke's theory of proper names, we have a neat explanation of all this: Oscar refers to Aristotle, because there is a causal chain leading from some original baptism of Aristotle to his use of the name; whereas Twin Oscar refers to Twin Aristotle, since the causal chain leading to *his* use is leading from some original baptism of *Twin Aristotle*.[12]) But whether someone accepts Kripke's theory of names or not, the important thing to remember is that as long as we find words whose reference is different when used by internally identical Oscar and Twin Oscar, the starting point of the externalist argument is secured.

Searching for more cases of this sort we may consider so-called indexical expressions like 'you' or 'she'. Suppose that Oscar has a friend called Lucinda, and Twin Oscar has a friend who is an exact replica of Lucinda. When Oscar and Twin Oscar simultaneously use the sentence 'You are beautiful' speaking to their respective friends, Oscar refers to Lucinda, and Twin Oscar refers to Twin Lucinda. That the references are different is as obvious in this case as in the case of proper names. So can we run a Twin Earth argument with 'you'? As we have seen, after establishing the difference of reference for our internally identical subjects, the next step in a Twin Earth argument is to use the connection between meaning and reference to draw a conclusion about meaning. However, we encounter a problem here. Putnam claims that in the case of indexicals, meaning does not determine reference.

Here are some reasons for this view. Ambiguity is the phenomenon when the same word has different meanings in a language; 'ball' could mean a festive event, or the

[10] Putnam 1970, 1975 and Kripke 1972. For more details, see the entry on NAMES AND NATURAL KIND TERMS.

[11] For a criticism of externalism along these lines, see Mellor 1977. See also the contributions on natural kinds in Pessin and Goldberg 1996.

[12] See the entry on NAMES AND NATURAL KIND TERMS.

round object used in a football game. This is a clear sense in which a word has different meanings on different occasions. However, when we use the word 'you' in different contexts to refer to different people, it is not ambiguous in this way. On the contrary, the natural assumption is that the meaning of 'you' is the same in all its uses, and presumably this is what we learn when we learn the meaning of the word. In the discussion above, we assumed that the claim that meaning determines reference amounts to claiming that there is a determinate reference belonging to every meaning, and consequently, difference of reference implies difference in meaning. In the case of 'you', this clearly does not hold: 'you' could refer to different individuals on different occasions and yet, as we just agreed, it has a constant linguistic meaning.

So far we have seen that the externalist argument has contradicted some of our initial assumptions about meaning. Could we reconcile the present finding with our reasons for holding that meaning determines reference? Let us consider a case when someone uses the words 'I' and 'you' *in the same context*: for example Oscar saying to Lucinda 'I stand by you'. 'I' in this case refers to Oscar, 'you' refers to Lucinda, and this difference is due their different meanings. Thus we could preserve the force of the original argument if we said that the meaning of an indexical determines its reference *within a context*, or with respect to a certain context. Next we should see what consequences this has for the externalist argument.

4. Sense determines reference

One of the first expressions of the idea that meaning should determine reference is found in Frege's famous paper, "On Sense and Reference": "to the sign there corresponds a definite sense and to that in turn a definite reference"[13]. Frege extended the sense/reference distinction to sentences; the sense of a sentence is a *thought*, and the reference of a sentence is its truth-value. The determination between sense and reference is upheld in the case of sentences: it is not only that thoughts *are* true or false, but also every thought has a fixed truth-value. How seriously Frege took this is illustrated by the way he deals with an apparent counterexample in a later paper, "The Thought". He notes that the sentence "This tree is covered with green leaves" may be true now, but false in six months' time. But instead of concluding from this that thoughts do not have fixed truth-values after all, he chooses to hold that the sentence expresses two different thoughts on the two occasions.

> The words "this tree is covered with green leaves" are not sufficient by themselves for the utterance, the time of utterance is involved as well. Without the time-indication this gives we have no complete thought, i.e. no thought at all. But this thought, if it is true, is true not only today or tomorrow but timelessly.[14]

[13] Frege 1892: 25
[14] Frege 1918: 103.

Some properties of thoughts may change – for example the property of being grasped by me or by someone else –, but the truth-value of a thought cannot. This suggests that according to Frege, *a thought has its truth-value essentially*. If two sentences differ in truth-value, they cannot express the same thought. The claim that thoughts determine their truth-value is an instance of the doctrine that sense determines reference. This, we can see now, is quite literally true: sense *alone* determines reference.

Closer inspection reveals, however, that this cannot be generally true. Consider a contingent sentence like

(1) The inventor of bifocals was a man.

As it happens, the description picks out Benjamin Franklin, who was indeed a man. So the sentence is true. If this statement is contingent, then there is another world where, say, Deborah Franklin invents bifocals, and where the sentence is false. Here we have a phenomenon which, at first sight, is similar to the one we encountered above about the tree and green leaves: the same sentence can be true in some circumstances (in our world) and false in others (in another possible worlds), so there is no determinate truth-value belonging to this sentence; an apparent counterexample to the claim that sense determines reference.

As a response, we could follow the Fregean recipe to the letter, and insist on the determination between sense and reference. This would mean that sentence (1) *expresses different thoughts in different worlds*. In other words, if we accept without qualification that thoughts have their truth-values essentially, then given that (1) is true, it is impossible to express the same thought by a false sentence. Hence in a world where Mrs Franklin and not Mr Franklin invented bifocals, the sentence could not express the same thought.

But contrary to this, it is standard to assume that in a world where (1) is false, its meaning or its sense or its content would nonetheless be the same. If this is right, then we cannot in general say that sense (or content) *alone* determines a truth-value; we also need the state of the world; that is, in this case, the fact that the inventor bifocals was a man. (An analogous reasoning holds for the description "the inventor of bifocals": its sense *alone* is not sufficient to determine its denotation. We also need the world to make its contribution.) When we say that sense determines reference, we understand this as relative to some state of the world.[15]

There is no immediate objection to extending the same strategy to indexicals: the suggestion would be that in the case of indexicals, meaning determines reference not only

[15] Dummett says that when we say that sense determines reference, "this is on the assumption that the contribution of extra-linguistic reality is being taken into account" (Dummett 1975: 121). But this is really an expression of the idea that sense *alone* does *not* determine reference.

relative to the state of the world, but also relative to a context. In the case of contingent (non-indexical) sentences, difference in truth-value *within a world* implies difference of content; in the case of indexical sentences, difference in truth-value *within a context* implies difference in content. Frege held that thoughts have their *truth-value* essentially; but today, we say that contents (which inherit the role of Fregean thoughts) have their *truth-conditions* essentially. This move is motivated precisely by considerations about sentences like 'The inventor of bifocals': for we can say that the truth-value of the sentence may vary from world to world, but the truth-conditions, and hence the content, remain the same. There is nothing inherent in the notion of truth-conditions which would forbid to say that analogously, though the truth-value of an indexical sentence may vary from context to context, its truth-conditions remain the same. Truth-conditions do not have trivial individuation in the way truth-values do. After all, we could have said that the truth-conditions of 'The inventor of bifocals was a man' are different in this world and in the other one: its truth depends on Benjamin Franklin's gender in this world, and on Deborah Franklin's gender in the other. Nonetheless, we decided to regard this difference as not affecting truth-conditions.[16] Why not make similar decisions in other cases?

The consequence of this is that the standard Twin Earth argument for externalism is inconclusive. Everyone should agree that meaning determines reference only together with some factors which are not themselves constitutive of meaning. It requires a separate argument to show that the context in which a sentence is used is not among these further factors. In the absence of such argument, it is possible to hold *both* that the meaning of an indexical determines its reference (relative to some further factors), and that it is the same for internally identical subjects.

Concluding his discussion of natural kind terms, Putnam says that 'Our theory can be summarized as saying that words like "water" have an unnoticed indexical component'[17]. Now even if Putnam's theory of natural kind terms is generally on the right lines, these terms do not function entirely analogously to indexicals. The reference of an indexical depends on the context of its *use*: if I travel from Budapest to London, the reference of 'here' shifts from Budapest to London. But if Oscar traveled to Twin Earth, the reference of his term 'water' would not – or at least not immediately – shift to XYZ[18] (and similarly, the reference of his term 'Aristotle' would not shift either). In the case of natural

[16] Compare here the fact that the premise used in the externalist argument is often formulated by claiming that meaning determines *extension*, and that difference in extension implies difference in meaning. This is clearly not right: even if the extension of 'philosopher' were different, this would be no reason to think that it had a different meaning. The most we would say is that meaning determines extension-*conditions* (analogously to truth-conditions). If 'reference' is the extension of a denoting expression, then meaning determines reference-conditions, and not reference. And the individuation of reference-conditions is also far from trivial.

[17] Putnam 1975: 234.

[18] Though it may shift after a certain time. This is the phenomenon called 'slow-switching", see Burge 1988.

kind terms, what seems to matter is not the context of *use*, but the context of *acquisition*. Still, keeping in mind these differences, we could extend the previous treatment of indexicals to natural kind terms and names. We could for instance say that the meaning of 'water' is the same for Oscar and Twin Oscar, and that this meaning – together with some further factors, like features of the environment where they acquire the word, or causal chains between initial baptisms and use of terms – determines reference. Thus contrary to the conclusion of the classic Twin Earth argument, the two assumptions that meaning determines reference, and that meaning is internal, are not incompatible.

5. External and internal semantic features

Everyone agrees that meaning determines reference only together with some further factors. Thus the idea that meaning determines reference, plus a mere difference in reference are not sufficient for the conclusion that meanings are different – the difference could be due to a difference in the further factors. There are, however, independent considerations which may show that in some cases where the reference is different, so are the contents of sentences. Suppose that you and I run a race; we hit the finish pretty much at the same time, and we both exclaim: 'You lost'[19]. We disagree; and this disagreement is naturally understood as stating different things or having different beliefs. A further plausible explanation of this is that the content of our statements and those of our beliefs are different. Suppose Oscar travels to Twin Earth, and pointing to a glass of water, says 'That's water'. If Twin Oscar says the same, and we acknowledge that their terms refer to different things, then again, it seems they disagree, and the disagreement is straightforwardly explained as having different beliefs, and that, in turn, that their beliefs have different contents. If someone wants to hold that the contents of these beliefs are nonetheless the same, something more complex has to be said about the semantics of belief attribution.

In the first kind of situation, it is natural to say that in some sense, you and I say the same, and in some sense, we say something different. The sentence "I thought I won; and she believed the same" is ambiguous: it allows us to assume that she thought that she won, or that she thought that I won. Two dimensional semantic treatments of indexicals try to capture this phenomenon by attributing two semantic features to indexical expressions: the first is constant throughout different uses, and hence makes the same contribution to the meaning of indexical sentences in every context. As we saw, this feature can be naturally regarded as the linguistic meaning of the indexical. The second feature may vary from context to context, depending on the actual reference of the indexical, making different contributions to the content of an indexical sentence. A further plausible thought concerns the relation between the two features: it is that the function of the constant meaning is to assign different contents to indexical sentences in

[19] I owe this example to Zoltán Gendler Szabó.

different contexts.[20]

The relevance of this to the debate about externalism is that while the utterances of indexical sentences by internally identical subjects share the first feature, they may differ in the second feature. There are many details of this debate which are discussed elsewhere in this book[21], we shall mention only a couple of points. First, can the two-dimensional treatment be extended to natural kind terms and names? While in the case of indexicals, it seemed plausible that uses of 'You lost' in different contexts are in some way similar, and some way different, and that *both* of these features are semantically important, the same is less obvious for names and natural kind terms. Suppose that someone holds the direct reference theory for names. Oscar and Twin Oscar have the same internal *physical* states; the symbols used by them may have the same *syntactic* features; but there is no *semantic* 'common factor' that their use of the name 'Aristotle' share. The only semantic feature of the name, its reference, is different. Of course, one can reject the direct reference theory for both names and natural kind terms, but the question remains: is there anything shared by say, Oscar's and Twin Oscar's use of 'water' beyond physical or syntactic features – something that has to do with *meanings*?[22]

6. The transparency of meaning

One reason why some philosophers have thought that there *is* an internal component to meaning is that certain features, which have been traditionally assigned to meanings, are apparently played by internally, and not by externally individuated aspects. One such feature is that meanings are known in a special way. [23] Michael Dummett writes

> It is an undeniable feature of the notion of meaning – obscure as that notion is – that meaning is *transparent* in the sense that, if someone attaches a meaning to each of two words, he must know whether these meanings are the same.[24]

[20] A classic treatment is Kaplan 1977: he calls the first feature 'character', the second 'content'. Though the details vary a lot, an important, and to some extent similar reaction to externalism about mental content was to distinguish 'narrow' and 'broad' (or 'wide') mental contents. McGinn 1982 and Fodor 1987 are classic versions.

[21] See the entry on INDEXICALS and on TWO DIMENSIONAL SEMANTICS.

[22] Stalnaker 1995 argues (as a response to Loar 1988) that there is no procedure which will result in a determinate, internally individuated content. A similar argument is in section 10 of Block and Stalnaker 1999. An argument against the view that mental content has an internal (as well as an external) component is in McDowell 1986.

[23] The questions raised by knowledge of mental contents and knowledge of meanings are largely the same; and the issue of externalism and the knowledge of mental content have inspired a very complex discussion, which I cannot hope to reproduce here. Some classic pieces are reprinted in Ludlow and Martin 1998. Wright, Smith and MacDonald 2000 includes further developments. A recent collection with contributions from many influential participants of the debate is Nuccetelli 2003.

[24] Dummett 1975: 131. Quoted in Boghossian 1995: 33. As Boghossian notes, the surrounding discussion makes it clear that knowing whether meanings are the same or not should be non-empirical.

Knowing which of my words have the same or different meanings teaches me that I can express my desire for sparkling water by using the words 'sparkling water' or 'fizzy water', but not with 'still water'. Of course such endeavours may fail if I grasp some meanings incompletely or incorrectly – that is, if I mean something else by a word than everyone else does. But it is a common assumption in the externalist arguments that when meaning is externally individuated, this is not simply the widely agreed phenomenon, discussed in the first paragraph of this paper, of incomplete or mistaken understanding. If Oscar doesn't mean *elm* by 'elm', and *beech* by 'beech', but instead he means some idiosyncratic concept *elch*, which has both elms and beaches in its extension, no externalist conclusion follows.[25]

Let us consider indexicals first – it was agreed that they have both an internally and an externally individuated semantic feature. I can tell whether two indexicals have the same linguistic meaning or not, and this will guide me in my use of 'you' instead of 'I', or of 'now' instead of 'tomorrow', as the situation requires. In contrast, the content – the externally individuated feature – of an indexical sentence depends on the context of use, and if two contexts are indistinguishable, I may not be able to tell whether the content is the same or not. 'You are one minute older than your twin' expresses different contents when addressed to Castor or to Pollux, but I may not be able to tell the difference; 'turning left here leads out of the labyrinth' expresses different contents when uttered at two different locations, but if the locations are indiscriminable, again I may be unable to tell the difference. Internally individuated features are transparent, while externally individuated features are not.

The phenomenon under discussion is the ability to tell whether two meanings are the same or not. This is stronger than the requirement of being able to tell, in some sense, what the meaning of a word is. Externalist theories of mental content were criticised on the ground that they cannot account for direct and non-empirical knowledge of our own thoughts. One standard response is the following. Beliefs about the content of our thoughts (e.g. the thought 'water is wet') arise from forming second-order thoughts ('now I am thinking that water is wet'). Since the content of the second order thought inherits the content of the first-order thought, there is no possibility of mismatch between the content of these two thoughts. Oscar and Twin Oscar start out with different contents for their first-order thoughts, and the difference is inherited to their second-order thoughts, so they both will be right. The contextually self-verifying character of these second-order thoughts is sufficient to explain the direct and non-empirical character of self-knowledge. As Burge says: "We 'individuate' our thoughts, or discriminate them from others, by thinking those, and not the others, self-ascriptively. … Our epistemic right rests on this immediacy …"[26]

[25] For an argument of how the externalist argument can be refuted if we allow such cases to count as simply misunderstanding or incomplete understanding, see Crane 1991.

[26] Burge 1988: 656. See also Burge 1996. Applied to meanings, the theory could be something like this:

Burge points out, correctly, that in order to know the content of our thoughts, we do not need to know every empirical fact which makes possible to have these thoughts. But even if we agree with this, it still may be objected that this theory provides a rather etiolated conception of self-knowledge, as an analogy will help to illustrate. I am always right in believing that *I am here*; a token of 'I am here' is contextually self-verifying in a similar way as second-order thoughts are. But even though I know I am here, I still may have no idea where I am.[27] It would not be particularly convincing to say that we individuate our locations, or discriminate them from others, by simply being at those locations, and not at others. This suggests that the impossibility of error because of the contextually self-verifying character of a judgement may not be sufficient to exclude ignorance. One way to spell out the idea that knowledge of my whereabouts is more than knowing that I am here, is to point out that what I lack in the case of not knowing my whereabouts, is an ability to discriminate between my present location and other locations. Analogously, a more robust knowledge of meanings and contents would require an ability to discriminate among them, and this is what Dummett's transparency thesis requires. This kind of knowledge is called 'dicriminatory' or 'comparative' knowledge of content.

On some externalist theories, there can be cases when we are unable to tell that two meanings are the *same*: on direct reference theories, for example, 'Hesperus' and 'Phosphorus' have the same meaning, but a perfectly competent user of the names may have no idea about this.[28] Are there cases where we cannot tell that the meanings of two words are *different*? Oscar and Twin Oscar's case poses no such problem: here the externally individuated meanings are possessed by two different language users. What we need is a single subject who can or cannot discriminate among different meanings. There are two ways to turn the Twin Earth scenario into such a situation. First, instead of imagining Earth and Twin Earth as two planets in the actual universe, we could conceive Twin Earth as a counterfactual scenario about Earth, and Twin Oscar as a counterfactual counterpart of Oscar. In the counterfactual situation, Oscar's word 'water' would have a different meaning. In the case of indexicals, we said that the same indexical sentence may express different contents in different contexts, but if the situations are indistinguishable, the subject may not be able to tell the difference. This applies to the counterfactual scenario we are considering: if Oscar had grown up on Twin Earth instead of Earth, his

---

both Oscar and Twin Oscar are right when they say "I mean *water* by 'water'". It is of course required that the sentences expressing knowledge of meanings or contents should be understood; I do not express knowledge by saying that '"Cantankerous" means cantankerous' if I don't understand what 'cantankerous' means. One way to put the issue between externalists and internalists is to ask whether there is a danger that on the theory just presented, knowledge of meanings reduces to people's 'mouthpiecing' such statements without really understanding them.

[27] This is not about the content of the thought 'I am here'; the point is an analogy, between knowing our location and knowing our thoughts.

[28] A similar case is Kripke's famous Pierre, see Kripke 1979.

situation would be indistinguishable, and hence by the same reasoning it seems that he would not be able to tell the difference in his concepts.[29]

Alternatively, we could try to furnish an actual language user with both concepts. Suppose that Earth and Twin Earth are part of the actual universe, and Oscar is transported to Twin Earth, unaware, overnight. When he wakes up, nothing seems different, and he goes on with his life. The general view is that after a certain time, the meaning of his word 'water', as describing his ongoing experiences, switches to twater. But there is no reason to assume that the concept that figures in his memories of Earthly water experiences switches too: when he recalls swimming in the Pacific back on Earth, and says that 'The water was salty', his word refers to H_2O. Similar phenomenon arises about his word 'Aristotle': when he remembers having read Aristotle's *Categories* ten years before, he refers to Earthly Aristotle; when he discusses his recent encounter with Aristotle's *Metaphysics*, he refers to Twin Aristotle. If externalist theories are right, the meaning of 'Aristotle' is different on the two occasions; yet Oscar is in no position to find this out merely by introspection (and similar remarks apply to 'water').

Inability to tell that two meanings are the *same* need not to be a consequence of all externalist theories; a view which combines externalism with Fregean senses[30] could hold that 'Hesperus' and 'Phosphorus' have indeed two different, object-dependent senses, and competent users of the name will be able to establish this merely by introspection. But the inability of telling that two meanings are *different* in certain situations – as illustrated by the slow-switching cases – is a consequence of all externalist theories. To see this, you merely have to assume that a single subject can be exposed to two different environments, and interactions with these environments result in acquiring different meanings. The differences in the environment, and the resulting difference in meanings cannot be traced to internal differences – otherwise we would not have a case of externalism. Hence the subject will not be able to tell merely by reflection the difference between these concepts. In contrast, an internalist will allow a difference in meanings only as long as it is traceable to internal differences.

It seems that internalism can provide a more robust account of knowledge of meaning than externalism can. But is there any reason to prefer the more robust account? Consider again the case where Oscar has been transported to Twin Earth, and his reports about his past experiences refer to Aristotle and water, the ones about his recent experiences to Twin Aristotle and twater. Oscar cannot discriminate between the two meanings of his word 'water', but he can discriminate the meaning of 'water' from the meaning of 'blood' or 'brandy'. The externalist then could say that this is sufficient to

[29] Burge uses counterfactual situations to set up his Twin type thought experiments for example in Burge 1979. In Burge 1988, he agrees that a person could not tell the difference between the actual and the counterfactual situation, but since he thinks that we could have knowledge of our thoughts even without being able to discriminate them from possible thoughts we might be thinking instead of them, he does not regard this as a problem.

[30] See Evans 1982, McDowell 1984.

award Oscar a knowledge of meanings. However, the internalist will have an objection. Suppose that Oscar argues in the following way:

1. "Aristotle doesn't refer to the notions of form and matter in his definition of substance in the *Categories.*"
2. "Aristotle's discussion of substance relies on the notions of form and matter in the *Metaphysics."*
3. "Therefore Aristotle has changed his views about substance between writing the *Categories* and the *Metaphysics.*"

This argument is mistaken; the question is, what sort of mistake is being made here. The internalist could say that it is a *factual* mistake: Oscar is wrong about the fact that the two books he read were written by the same person. But on the externalist view, though Oscar may make a factual mistake, he also makes a *logical* mistake: he equivocates on the word 'Aristotle'; in the first and the second premise, the word has different meanings. Of course we do make logical mistakes for example when a subject matter is complicated, when we don't quite understand the concepts, when we are in a hurry, when our judgement is discoloured by emotions, and so on. But even though Oscar is completely dispassionate on this topic, he is a perfectly competent user of all these words, the whole issue is quite simple, he has all the time in the world – he still won't be able find out this logical mistake simply by reflecting on the premises of his argument. [31]

Faced with this situation, we could simply draw the consequence that the kind of mistakes we are prone to in our empirical or perceptual judgements may affect also our judgements about meanings or mental contents. My rationality is not threatened if I cannot tell Castor and Pollux apart just by looking; and similarly, this argument continues, it's entirely understandable if I cannot discriminate some of my meanings introspectively.[32] However, one might want to distinguish between factual and logical mistakes of this kind.[33] And if someone thinks that therefore the consequences of the externalist view pose a serious threat to our rationality, she should object to the view which entailed it.[34]

[31] For other cases of logical mistakes, entailed by an externalist conception of mental content, see Boghossian 1994.
[32] See Owens 1989.
[33] I have in mind the kind of distinction drawn in McDowell 1995 (see esp. fn 5); mistakes which are, or not, results of 'misconducting oneself in the space of reasons'. Of course, McDowell would not subscribe to this point being used in an argument against externalism.
[34] I am grateful for Tim Crane and Zoltán Gendler Szabó for discussions and comments, as well as for the support of the Hungarian OTKA (grant no. T046757) and the Philosophy of Language Research Group of

References

Block, Ned and Robert Stalnaker 1999. “Conceptual analysis, dualism, and the explanatory gap.” *Philosophical Review* 108: 1-46.

Boghossian, Paul A. 1994, "The transparency of mental content" *Philosophical Perspectives* 8: 33-50

Burge, Tyler 1979: "Individualism and the Mental" . Reprinted in David M. Rosenthal (ed.): *The Nature of Mind* Oxford University Press 1991: 536-567 and in Ludlow and Martin 1998: 21-82. Some sections are reprinted in Pessin and Goldberg 1996: 125-141.

Burge, Tyler 1982: "Other Bodies" in: Andrew Woodfield (ed.): *Thought and Object*. Oxford: Clarendon Press. Reprinted in Pessin and Goldberg 1996, 142-160

Burge, Tyler 1988: "Individualism and Self-Knowledge" *Journal of Philosophy* 85: 649-63. Reprinted in Pessin and Goldberg 1996: 342-354, and in Ludlow and Martin, 1998: 111-128

Burge, Tyler 1996: "Our Entitlement to Self-Knowledge" *Proceedings of the Aristotelian Society* XCVI: 91-116. Reprinted in Ludlow and Martin, 1998: 239-264

Crane, Tim 1991: "All the Difference in the World" *The Philosophical Quarterly* 41/162: 1-25, Reprinted in Pessin and Goldberg 1995, 284-304

Dummett, Michael 1975, “Frege’s Distinction between Sense and Reference” in *Truth and Other Enigmas*. London: Duckworth 1978: 116-144.

Evans, Gareth 1982, *The Varieties of Reference*. Oxford: Clarendon Press.

Farkas, Katalin 2003, “What is externalism?” *Philosophical Studies* 112/3: 187-208

Fodor, Jerry A. 1987: *Psychosemantics*. Cambridge, Mass.: MIT Press

Frege, Gottlob 1892, “On Sense and Reference”. Reprinted in A.W. Moore, ed. *Meaning and Reference*. Oxford University Press 1993: 23-42

Frege, Gottlob 1918, “The Thought”. Reprinted in Simon Blackburn & Keith Simmons, eds., *Truth*. Oxford University Press 1999: pp. 85-105

Kaplan, David 1977: "Demonstratives" in: Almog, J. & Perry, J. & Wettstein, H. (eds.) 1989: *Themes from Kaplan*. Oxford University Press: 481-563

Kripke, Saul A. 1972. ‘Naming and Necessity’ Reprinted in book form by Blackwell, Oxford 1980.

Kripke, Saul A. 1979: "A Puzzle About Belief" Reprinted in Salmon, Nathan & Soames, Scott (eds.) 1988: *Propositions and Attitudes.* Oxford University Press:102-48

Loar, Brian 1988. “Social content and psychological content” Reprinted in Pessin and Goldberg: 180-192

Ludlow, Peter and Norah Martin (eds.) 1998, *Externalism and Self-Knowledge* Stanford: CSLI

McGinn, Colin 1982, “The Structure of Content” in Andrew Woodfield (ed.) *Thought and object*. Oxford: Clarendon Press: 207-58

McDowell, John 1984 "*De Re* Senses." *Philosophical Quarterly* 34(136): 283-294.

---

the Hungarian Academy of Sciences.

McDowell, John 1986: "Singular Thought and the Extent of Inner Space" in Pettit, Philip and John McDowell (eds.): *Subject, Thought and Context* Oxford: Clarendon Press: 136-168

McDowell, John 1992: „Putnam on Mind and Meaning." *Philosophical Topics* 20(1):35-48. reprinted in Pessin and Goldberg: 305-317

McDowell, John 1995 Knowledge and the Internal," *Philosophy and Phenomenological Research* 55: 877-893

Mellor, D. H. 1977: "Natural Kinds" *British Journal for the Philosophy of Science* 28: 299-312. Reprinted in Pessin and Goldberg: 69-80

Newman, Anthony *forthcoming*: "Two Grades of Internalism (Pass and Fail)" *Philosophical Studies*

Nucetelli, Susana ed. 2003, *New Essays on Semantic Externalism and Self-Knowledge*. Cambridge, Mass.: MIT Press. Bradford Books.

Owens, Joseph 1989. "Contradictory Belief and Cognitive Access" Midwest Studies in Philosophy, XIV, University of Minnesota Press: 289-316

Pessin, Andrew and Sanford Goldberg (eds.) 1996, *The Twin Earth Chronicles* Armonk, NY and London: M. E. Sharpe

Putnam, Hilary 1970. "Is Semantics Possible?" Reprinted in *Mind, Language and Reality*. Cambridge University Press 1975: 139-52

Putnam, Hilary 1975. "The Meaning of 'Meaning'" in *Mind, Language and Reality* Cambridge University Press. Reprinted in Pessin and Goldberg 1996: 3-52.

Searle, John R. 1983. *Intentionality*. Cambridge University Press

Stalnaker, Robert 1995. "Narrow Content" Reprinted in *Context and Content* Oxford: Clarendon Press 1999 p. 195-210

Strawson, Peter F. 1959: *Individuals.* reprinted in 1993, London and New York: Routledge

Williamson, Timothy 2000: *Knowledge and its Limits* Oxford University Press

Wright, Crispin, Barry C. Smith, and Cynthia Macdonald (eds.) 2000**,** *Knowing our own minds*. Oxford : Clarendon Press