Putting One's Foot in One's Head —Part I: Why

Donald Perlis
Department of Computer Science
and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
(301) 405-2685
perlis@cs.umd.edu

Abstract

The studies of mind and language have traditionally been linked to one another. Indeed, theories of reference (meaning, intentionality, content) have over time brought more and more mind into meaning. Here I argue that the links must be made far stronger still if we are to understand either. I offer some criticism of the causo-functionalist theories of reference on this ground, and present some ideas for improvements. The upshot will be that intentionality is largely internal and very real indeed, that it provides a genuine distinction between systems that have it and ones that do not, and that it hinges on internal complexities of a specifiable sort, ones that are tied to the detection of error.

What is meaning, what is it good for, and what is a naturalist (non-intentional) account of how it works? In this paper (part I of a two-part work), I will argue that the first two of these questions should be answered primarily in terms of processes inside the mind/brain, contrary to most recent thinking on this. In the sequel paper (part II) I will argue that a key aspect of this is the physical body of the meaning agent, and propose a way in which this might work, and how an external notion of reference can be recaptured from internal processes.

This is a preprint of a paper to appear in Noûs.

1 Introduction

What makes my stick figure an image of a tiger is not that it looks much like one...but rather that it's my image, so I'm the one who gets to say what it's an image of. My images...connect with my intentions in a certain way; I take them as tiger-pictures for purposes of whatever task I happen to have at hand.—Jerry Fodor, *Language of Thought*, p. 191.

Fodor was here defending imagism against Dennett's onslaught, but I employ the quote for a different reason: note the word "intentions". People intend things by their words and actions and thoughts. It may be no accident that Brentano borrowed the term "intentionality" from the scholastics; it may turn out to have everything to do with people's intentions after all; and so I will argue.

In this paper (part I of a two-part work) I present arguments why it is essential to look deeper for internal mechanisms pertaining to a reference relation. I will do this largely by considering the highly successful causal-history theory of reference, and suggesting particular points at which it seems to need modification or extension. This will focus on the key aspect of grounding or dubbing that plays a dominant role in that theory. I will argue that dubbing is largely an internal matter. Finally I try to show that an internal perspective has benefits in terms of clarifying the advantages that language and thought confer on intelligent agents. In the sequel paper (part II, Perlis, forthcoming-b) I will provide an outline of a theory along the internalist lines called for in the present paper, and I will discuss it in connection with various matters including Fodor's (1987,90) theory of asymmetric dependence. The present approach has similarities to (Minsky 1968, Perlis 1987, Rapaport 1988), and should be seen as attempting to extend such ideas and compare and combine them with alternative approaches, especially in the philosophic literature.

I will start by reviewing the problem (section 2) and in section 3 a bit of the history of thinking on it, leading to a particularly interesting approach—the causal-histories theory of Kripke (1980) and Putnam (1975). In section 4, I suggest some ways in which this causal theory needs to be modified, along lines of greater internalism (already hinted at by Kripke). In section 5 this is further discussed in terms of simple architectural models of representation. We return to a concrete example in section 6, where dubbing is argued to be a central internal feature of reference in general, and this will lead us to a notion of "input verificationism" in section 7. Finally in section 8 we discuss advantages of this view. A treatment of how internal-dubbing bears on external reference will be left to the sequel paper. Although the literature on the topic is vast, I will most often refer to the excellent survey text by Devitt and Sterelny (1987), especially as they espouse and elaborate on the causal theory that I am both criticizing and trying to improve.

2 Inner talk

All our thinking goes on in the head, with head-stuff (language, symbols, neurons, whatever). If in our thoughts we take "foot" to have a meaning, that taking goes on in our heads. Yet there's no foot in our heads, so what has our thinking "foot" or taking "foot" to mean something, to do with an actual foot out there? How do inner manipulations ever get related to external things?

This is in part Descartes' observation, that there is a Divide between what we think and what is true, for our minds can be deceived in any number of ways. Our minds are not in direct contact with things out there. But there is another issue, raised by Brentano (1874): It is not simply that our thoughts may be false, but that they may have no meaning. In the absence of a link between internal and external, what can give the internal manipulations meaning? If there is no evident link between our mental tokenings of the word "foot" and a real physical limb out there, then in what sense does the word mean the limb? What is it that provides the link, the directedness, between the (internal) word "foot" and my (external) foot?

Answers suggest themselves, of course. But close analysis has shown the easy answers to be unsatisfactory in various ways. One problematic aspect has to do with terminology. If our words and thoughts cannot get outside our heads, how can we even talk about meaning? We seem reduced to "inner talk", yet we appear on the surface to be talking about all sorts of outer things.

This is a problem that philosophers from Berkeley on have struggled with, and that has concerned the recent Putnam (1981,1987). Husserl (Brentano's student) introduced useful terminology: he speaks of "bracketed" objects of belief (see Magee, 1988, p. 256; and Putnam 1981, p. 28). Thus person P, in seeing her dog Sandy, first and foremost sees what we may write as [Sandy], i.e., this is her visual experience, what she takes to be Sandy. This is much like Dennett's "notional world" (Dennett 1987, pp. 152ff) of the meaner, or Devitt's (1989) "proto-reference", or what I will sometimes call one's world view or mental space (not to be confused with the more technical use of the expression "mental space" in (Fauconnier 1985)). In P's world view there is a dog called "Sandy". She refers to it, in a bracketed sense, whether or not she is deluded and has no dog. Of course, [Sandy] need not be a visual image. It is simply whatever it is (in P's head) that is what P takes to be called by the name "Sandy"—i.e., P takes "Sandy" to refer to [Sandy]. This raises conceptual difficulties; for instance, P probably would not agree that she takes "Sandy" to stand for something in her head, at least not the way that would usually be construed. P thinks of the referent of "Sandy" as out there. But in so thinking, P has set up in her head a something-or-other that plays the role of dog-Sandy, something in her world-view. And her world view is in her head.

This can be made more graphic with an example borrowed from Michael Ayers (see Magee 1988, p. 124): suppose "Durer's idea of a rhinoceros was not much like a rhinoceros." That is, Durer had it wrong as to what rhinos look like. Yet there is another sense in which the statement can be read, namely, that Durer's ideas are in his head, and there everything is brain tissue, not much at all like rhinos (except for rhino brain tissue!). In this sense, no one's idea of a rhino is much like a (whole) rhino. Of course, when Durer goes to paint a picture of a rhino, he paints it as he imagines rhinos to be, and the painting will not look much like a rhino (on the first reading of Ayers's statement). For instance, we might suppose for the sake of argument that Durer believed rhinos have no horns. Now once he paints a picture of a hornless animal, we can make easy sense of the picture not looking like a rhino, for rhinos have horns. But the idea in Durer's head is a hornless-animal idea—and how are we to compare ideas to rhinos? We cannot get our hands on ideas, and if we could get at the brain tissue (and associated electrochemical processes—let's call this whole mess DR) that constituted Durer's idea of a rhino, what would it mean to say it represented horned or hornless things?

What is presumably wanted here, to sort this out in a satisfying way, is some kind of means to identify invariants in brain process, so that my idea of a rhino can be compared to your idea, and even more basic: so that brain process can be mapped to one's *view.* The complexity DR in Durer's head is to *him* a rhino, i.e., it is his thinking of a rhino. What makes it so? It is utterly un-rhino-like in itself. I will be offering an account below, of how we may be able to assign *intrinsic* meaning to certain mental structures, in a way that may explain our intuitive sense that our ideas have meaning for us independent of what external expression of those ideas may convey to others.

To return to Sandy: the tough question, the standard problem of external reference, is what determines that it is in fact the real dog Sandy that [Sandy] links to, in the "usual" case in which there is such a supposed external referent (no delusion)? This external component of reference I take up in the later paper. For the present, I hope to have made a prima facie case for the need for a reference relation internal to the meaner: P must have both the symbol "Sandy" and its meaning as *she sees it*, in order for "Sandy" to be a symbol at all. And it is in her control: the choice as to what "Sandy" is to mean for her is arbitrary, in that it is governed by her, not by convention or form or entities outside her. P could just as well decide to rename her dog, for instance.

I will argue these points more fully later. For now, note that despite Harman (1975) and Devitt (1981, pp. 97–100) there is no infinite regress here. That is a customary attack on so-called competence theories of language. But I am not requiring that P understand or represent the symbol "Sandy" as well as the notional dog [Sandy]. There are cases in which such further representation is called for, and in these cases (such as the renaming alluded to above) another layer of symbolism is needed. But there is no claim here that such layers are present at all times. Nevertheless, I will claim below that for many purposes, we do need at least one further layer of representation, without which we are severely deprived of some basic commonsense behaviors. For now, P represents [Sandy] by "Sandy", and that is all.

3 Some history

Aristotle and Locke suggested that we associate words with ideas or images. These provide "immediate" meanings for the words. In turn, the ideas are related to external meanings by "mediation" of the senses (or brain). However, there remained a lack of cogent specification of what constitutes an idea or image, and how the mediation provides the needed sort of link in a general way (e.g., how the idea resembles the object of mediate reference). But at least this view provided an explanation of failure to refer: we can have, say, a Santa Claus idea, even though nothing external answers to that idea.

Brentano, as noted above, suggested that the directedness of word (or thought or emotion) to an external referent provided a characterization of the mental: nothing else seems to have this directedness. And he proposed a device to carry this directedness, even in cases of reference failure, namely, "intentional inexistents". Meinong (another student of Brentano) proposed that these objects of thought be the focus of the study of metaphysics.

Mill is sometimes (dis)credited with the "direct reference" view of semantics for proper names: that the only role a name plays is as stand-in for the thing it names. That this is problematic is easily seen: Hesperus and Phosphorous name the same thing (Venus) but they do not play the same roles in our thinking. Similarly for 1 + 1 and 2.

Frege and Russell proposed that descriptions or senses lead from word to referent, and thus that Hesperus, for instance, is associated with the description or sense of "evening star" and thus plays a distinct role from Phosphorus ("morning star").

It is interesting that the idea/imagist theory, the intentional-inexistent theory and the description theory all invoke a triune relationship, in which some intermediate entity E carries the semantic burden: the word is associated with E, and in turn E with the external referent (when reference succeeds). The description theory might seem to be the most robust version of the three, since it alone provides some technical hints of how the relation is determined.

However, apparent flaws were found in this sort of theory, notably by Putnam and Kripke. Consider the name "Albert Einstein". Perhaps we associate it with the description "the discoverer of the theory of relativity". But we might also associate with "theory of relativity" the description "the theory Albert Einstein discovered". If so, then we are not led out to either an actual theory or an actual person. Yet many of us would seem to know nothing more about Einstein and relativity than just this. So associated descriptions apparently do not account for meaning in general.

Putnam and Kripke suggested an alternative, the so-called causal-history approach. When a name comes into use, it is at first associated with its referent by those who coin or create the name, in a "grounding" or "dubbing" act. For instance, as in (Devitt and Sterelny 1987), one might decide to name one's cat "Nana". Later, others are informed of this dubbing (perhaps indirectly), and they too come to use "Nana" to mean that same cat, even if they have never seen the cat. Thus reference is borrowed again and again, and eventually people who have no idea what "Nana" refers to may still use that name meaningfully, by saying, for instance, "I heard that Nana is ill", and such an utterance may be true without the speaker having any way to verify this, except by asking others who may happen to know more about Nana. Thus over time a causal history of borrowings spreads the usage of "Nana" across a linguistic community, and allows us (in principle) to trace an utterance involving "Nana" back to its origins and thereby determine whether the utterance is true.

Thus causal chains of designation leading back to original dubbings serve to ground the meaning of the term by present users. In a way to be argued below, this may provide a handle on the resemblance feature of the Lockean view that between word and object there are ideas that provide the immediate meaning of the word and which in turn is linked to the mediate object of reference.

This ends the historical sketch. In the remainder of the present paper (Part I), I will focus on some inadequacies of current (causal) approaches, and in particular urge the need for a more robustly internalist view of meaning. Then in the sequel paper (Part II) I will attempt to carry out the development of such an internalist theory.

4 The need for more internalism

Let us again ask, What is meaning? What do we expect a theory of meaning to explain? Following Devitt and Sterelny (1987), it is sensible to look for the roles meaning plays in our lives, and what we hope to be able to explain by it. Two such roles stand out (Devitt and Sterelny say): learning about the world, and predicting our behavior. Now, there are two main "modes" of meaning that have been given attention in the recent literature: external or wide meaning or content or reference (involving truth conditions) and internal or narrow meaning or content or reference (in the brain, mind, or world view of the meaner). Causal theories of meaning (or reference or content) are aimed at the external notion of meaning (although Devitt argues that this includes the internal as a part of it). External meaning or reference (*eref* for short) seems well-suited to explain the *success or failure* of behavior, and also the "learning about the world" role of meaning.

On the other hand, internal meaning seems well-suited to explain behavior itself, i.e., why an agent chooses to do something. Internal meaning or reference (*iref* for short) has to do with how the agent sees the world, not how the world actually is. This puts iref closer in spirit to description theories of reference than to causal theories. For if we think of a description as information in the agent's head, concerning the use by that agent of a token in mentalese, then the agent may base her behavior on that information, and whether or not there is a similarity between that information and the external reality is a separate issue. For the same reason, iref has more to do with mind than has eref (e.g., as the Twin-Earth scenarios suggest). ln fact, we will suggest below that neither internal nor external accounts alone are sufficient to explain either truth conditions or behavior.

Causal theories would seem at first glance to provide a satisfactory account of eref. However, there are standard problems with such approaches, which I will list here with only brief explanation; some of these will be treated in the sequel paper.

1. The "qua" problem: Any explanation of external reference must give an account of how it is that the cattiness, and not the mammalhood, of that creature before us is what makes it the referent of "that cat"? How does "cat" get fixed just to cats and not a larger (or smaller) class of entities? This is a severe problem for causal history accounts of reference.

2. The "error" problem: Any explanation of external reference must give an account of how reference can go wrong. This problem is especially puzzling for a verificationist/reliabilist/covariational account of reference. If that which reliably occasions the usage of t is the meaning of t, then how can it happen that there are systematic errors in occasioning of usage?

3. The "false positive" problem: Any explanation of external reference must give an account of how harmless errors can occur without disrupting the very notion of reference being supplied. Some terms t seem to be used (harmlessly but) incorrectly more often than correctly, so what is it that makes the few correct usages really "correct"? This is much the same problem as the "error" problem, except that in this form it is particularly difficult for teleological/evolutionary/biological accounts of reference.

4. The "indeterminacy" problem: What makes the referent r of a term t that entire referent, all of r, and not part of it. Why does "that cat" mean the whole animal and not just the part I can see, or just that animal's life today (not yesterday or tomorrow)? What provides

the extent, the boundaries, of the referent? This is sometimes also considered to be part of the "qua" problem. As such it is again a difficulty for causal historical accounts.

Of these, 1–3 are discussed at length in (Devitt 1990), and 4 by Quine (1961) and Putnam (1981,1987). All of these, I think, remain problems because the causal theories (in their various versions, of which we have only here discussed the "historical" one) try to do too much on their own, without sufficient interaction with accounts of iref. Recently Devitt (1989) has written on this as well, but not with the same focus as the present paper: he regards iref as a thin version (abstraction or portion) of eref, whereas I regard them as importantly distinct. In short, all these problems seem to me to arise because the meaner's viewpoint has been given short shrift. In this paper I will be concerned with elaborating this outlook, and showing how a more robust account of meaning and mind can be given by paying more attention to internal subjective viewpoints. However, I am firmly on the side of naturalist and computational accounts of the mental; I am not trying to use mentalism as a foil against progress of the naturalist approach as Dennett (to appear) argues—incorrectly, I think—of Fodor. In fact, I will propose a naturalist, non-intentional solution to the problem of mental content.

External approaches are ones that try to locate meaning largely outside the head, in the things that serve as referents. This is of course an obvious thing to try to do. But I think it is wrong, and reveals a confusion as to what our goals are. Consider an ordinary speaker, who says "X" to mean Y. She says "X" because to *her* "X" means Y. And most likely the reason "X" means Y to her is that "X" means Y in her linguistic community, from which she picked that up. But what is it that she has picked up? This story says absolutely nothing about what the meaning relation is, but only that it is often shared among people. What is it for "X" to mean Y, for a person to mean Y by "X"? It tends to be somewhat uniform and spreadable within groups, but that is precious little to know about something. If all we knew about forest fires was that they tend to spread from one tree to another, and in so doing tended to be rather similar from tree to tree, we would know little at all about them. We would not know they were hot, dangerous, involving rapid oxidation, etc. Yet, it seems to me, most of the work on meaning (Fodor is a notable exception) has been on the meta-properties of meaning, not meaning simpliciter. Something goes on in a tree when it burns, more than is expressed by saying the fire has reached it; and similarly something goes on in a person when meaning reaches her, far more than is expressed by saying she has picked up a new word. We have all this stuff spread around, and we know something of how it spreads, but not much about what it is. I have left something out, though. Externalists also look at the relation between a word and its referent; this is after all the basic idea. But they do so via the spreading of the relation, rather than via a good hard look at the relation itself. In fact, they speak as if the spread was the relation, and that nothing inside the person matters very much. There are occasional nods toward the person herself, but it has not been explored very seriously. The approach seems to have more to do with etymology than with mind.

The above caricature is unfair, however. Many, such as Devitt and Putnam, not to mention Fodor, are indeed interested in more detail about meaning, but simply despair of finding anything uniform enough across instances of meaning, to make for a useful theory of the inner workings of a meaning agent. I will argue below that it is essential to go beyond the social network, and to look within: another internal component must be added to the functionalist cognitive architecture.

We can perhaps forestall an early objection: of course it is impossible to derive external content *solely* from internal mechanisms alone. This much I grant Putnam has shown, and Descartes before him. But the amount of external mechanism needed to cross the Cartesian Divide is less than one might have thought, or so I shall argue.

5 Representation reconsidered

Now we back up in light of the preceding, and study the basic question more abstractly. What is it then for something to be taken as a sign of something else? We follow our earlier discussion of P's dog Sandy. In general, for P to take a token S as sign for R, the taker P must have mental access to both S and R. That is, both S and R must in some sense be present in P's mental activity. We have indicated this earlier via bracket notation: [R] is P's notional version of R. In what follows we will more commonly use R_P instead of [R], to emphasize that it is person-dependent. Of course, there is some question-begging here: what does it mean to say that the notional object R_P corresponds in any way to an external object R? However, even without an answer to this question, there seems to be good reason to suppose that an internal reference relation (that between S and R_P) is a crucial part of the phenomenon of meaning. We have argued this a bit above, and present a more detailed argument below.

We can illustrate the point in various ways, starting with a much simpler account, which will then push us back to that above view. Suppose that P takes S to mean R. We suppose that S in fact is a token in P's head (e.g., the word "Sandy"). R is an external object supposedly pointed at by S for P. Yet P has no pointer to R! This I call passive representation. There is no activity or behavior of P that relates S and R. We, looking in, may relate S to R, but that is not something that P herself does. The significance of this will become clear as we proceed. S at the very best is a copy of R (e.g., a visual image), but even this does not serve to link S to R for P.

Passive representations will not do for an account of meaning. Consider an electric eye that opens a door when a person passes in front. Here P is the electric eye and associated door-opening mechanism; R is an object that passes in front of the door; and S is a signal that propagates inside P when R passes. We now may describe S as meaning or representing R's passing; or as representing the door being about to open; and yet again as the representing the blockage of light to P. All are relevant, none has special claims to our attention, except as regards our interests and intentions. There is no determination within P that singles out any special meaning among these.

Thus the notion that a single token, S, inside an entity P, can *represent* unambiguously a canonical external referent R, seems false and unhelpful. Yet to a large extent this is the model that is presented in studies of meaning and reference. The meaning is to be given principally in terms of external features. Putnam calls this model the copy model; I call it the one-tiered model. (See Figure 1.) It ascribes too little to the person P. In particular, it leaves out the key feature that P herself is doing an act of *taking* S to be something. It is not just that P *has* S, but that P takes S to stand in for something else. This means P must also have a means to relate, in her mind, both S and what she takes S to represent. But how does P do anything mental with respect to R, if R is outside P?

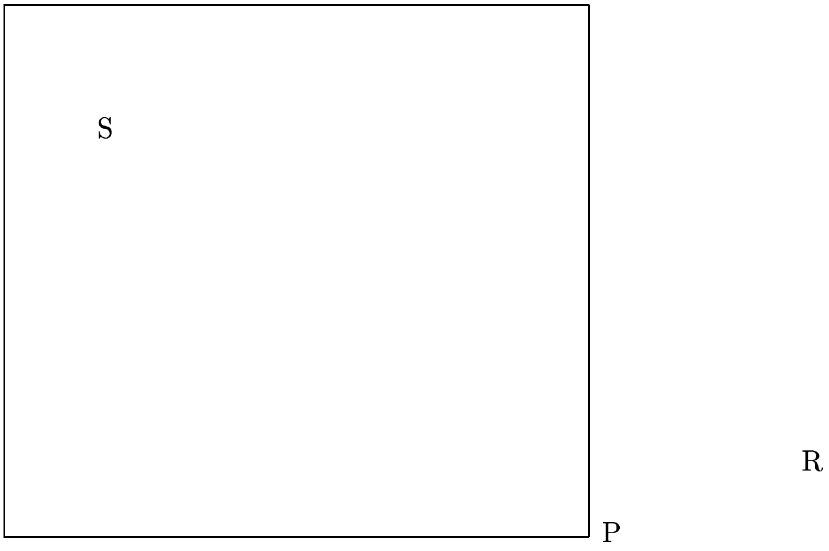

Figure 1: The One-Tiered Model. Here S is at best only a copy of R in P; no meaning is being conferred on S by P.

P must take herself (via internal means) to be referring to something supposedly out there. Now, this does not solve the problem, but it is a start, that we will try to develop into a solution in the rest of this paper. Let us see how it is a start. We postulate that in P's head there is not only S but also a description of S as being taken to be something out there. Let us use a name again, for more concreteness. Suppose S is the name "Sandy", and that P has in addition a visual image R_P of (i.e., causally produced by) P's dog Sandy, and also the information (description) that "Sandy" names R_P. That is, P takes "Sandy" to be the name of that doggish thing (R_P) in her visual field. Note that S (the word "Sandy") is not the notional dog; rather R_P is the notional dog. Such a "two-tiered" model is shown in Figure 2.

This I call active representation. P can manipulate S internally in a way that can affect its internal relation to R_P. The specific advantage will be further drawn out later. Roughly, it has to do with the ability to formulate alternative hypotheses concerning what is "out there".

Here is where things get tricky again along the lines of the rhino problem. Our ordinary terminology fails us, since we are used to saying P sees Sandy and not the image R_P of Sandy. But of course, what is going on when P "sees Sandy" (as we would normally say) is that there is an image R_P of Sandy in P's head, which P takes to be Sandy—unless of course P is in an introspective mood and reflecting on the behavior of her visual system, in which case she takes R_P to be an image of however she *then* represents Sandy in her mind; but her mind never manages to reach out of itself to the actual dog Sandy. Her "notional" Sandy is in her mind. Just how she uses that internal entity (i.e., what good it does her) and how it relates to the real external Sandy are the topics calling out for explication. Part of this we address below, and part in the sequel paper.

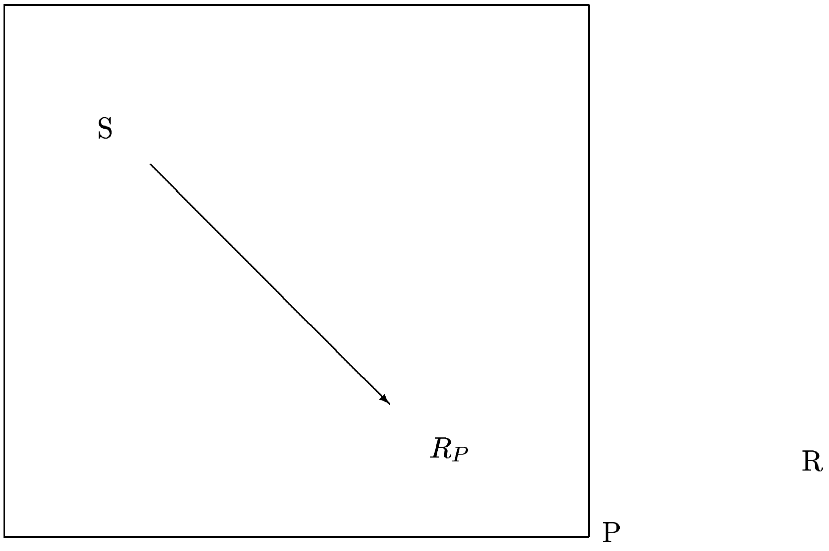

Figure 2: The Two-Tiered Model.

6 The dub's the thing

In an effort to come to grips with "public" meaning (and eref) in natural language utterances. the causal theory has done a great service. But it does so at the expense of internal "mental" meaning (iref). Devitt and Sterelny (1987, pp. 119–127), attempt to patch the causal-history theory with a Gricean theory along these lines. But they leave unexplored something vital to the enterprise. The entire causal-history theory itself rests on the notion of dubbings that provide the basis for the spread (borrowing) of meaning throughout a community. But both spreads and dubbings are profoundly *internal* events, as we now argue. In fact, we will argue that borrowings are dubbings.

Let us consider an example, modified from Devitt and Sterelny (p. 64). Sue overhears use of the word "Purdue" and from the context comes to believe Purdue is a person. Her actions are best explained in terms of the "person" meaning (i.e., her iref) for "Purdue". In a description theory reading, Sue has a description of what she takes to be the referent of "Purdue" which describes that referent as a person, not a university. In our earlier notation, Sue's S is the word "Purdue", and R_{Sue} is a person in her notional world (the person she takes to be Purdue, i.e., to be named "Purdue"), so the iref link goes from "Purdue" to $Purdue_{Sue}$. What the external R is here, is the crux of the matter. In fact, there may (or may not—see below) be two R's: Purdue University and some actual person that Sue may dub as the entity named "Purdue". On the standard causal-history reading, R is Purdue University.

The point I want to call attention to is that Sue has here in her own view of the world a representation of a reference relation: she *takes* "Purdue" to refer to a person. It is not simply that she has a disposition to act in such-and-such a way. There is in her beliefs the information that "Purdue" means someone-or-other. She might in fact believe that the man with a cane who passed by moments earlier is Purdue; that is, she might perform a dubbing act by so linking him to "Purdue"; in this case there are two external R's. Or, she might have no particular person in mind, but simply take others to be talking about someone or other named "Purdue"; in this case it is less clear whether there are two R's. In either case, however, she carries out some internal activity that forges a link in her mind, between "Purdue" and $Purdue_{Sue}$. It is this that makes "Purdue" a meaningful token for her. Indeed, her subsequent behavior will conform to the "person" meaning, and not the "university" meaning; upon being told that she will be taken to see Purdue, she will prepare herself for introductions, not for a walk across campus.

Moreover, Sue does not make the iref link completely on her own. She takes herself to be using it the same way the others are; she takes herself to be borrowing reference. She performs a sort of dubbing of her own, in deciding to use "Purdue" to mean such-and-such (in this case, to use it in what she supposes is a conventional way). Thus Kripke (1980, p. 96) says that the borrower "of the name must...intend to use it with the same reference" as the person from whom it is borrowed. It is not enough that the name itself be borrowed; the reference must be borrowed too, at least in intent. Let us call this Kripke's Condition. This means that the borrower must take the name to have a referential role: it is a term that refers to something, X, that the borrowee (the reference lender) had in mind. That is, the borrower must *intend* to be tied into the causal chain. This intent amounts to a description, indeed a prescription, for determining the term's meaning as the borrower intends it: it is "whatever was meant by the borrowee". This description is both a partial determiner of the (intended) meaning and also a possible barrier to its being known in detail to the borrower, for she may not know what the borrowee has in mind (and even less what the original dubbers had in mind). She has a description that tells her that her usage (as she herself defines it) involves parameters outside her. These parameters are ones that in principle she might be able to track down, but when she does they may surprise her: Purdue turns out to be not a person but a university, in the appropriate causal-chain.

Thus Sue's notional $Purdue_{Sue}$ is a conflation, for it simultaneously involves (say) "that man with a cane" and "what they are talking about", and these two do not in fact describe anything at all. But there is a uniform feature common to all borrowings: they involve an internal dubbing of some notional entity R_P by some token S. The same of course holds for original dubbings. Now, this means that notional entities such as R_P can be very complex: to be both a person and a "what they are talking about" is, at the very least, to be complicated. But I think this is essential: mental states with genuine content will be complex; our job is to look into this complexity, and determine its broad outlines. (Possibly the more technical use of the expression "mental space" in (Fauconnier 1985) will be helpful here.)

In this sense, then, we know what we mean by our words. We may not know what others mean by their words in detail, and we may even have it all wrong (what Sue means by "Purdue" is not at all what others mean) but we at least know what we mean. Otherwise we are merely intoning sounds when we speak.

We can relate this to Locke's (1689) resemblance criterion. When a term is borrowed, the borrower P takes it to mean whatever the borrowee Q meant. This provides an indirect (borrowed) resemblance between R_P and R, for (let us suppose) R_Q really does resemble R. However, this is still inadequate for many cases, such as reference of "proton" or "happiness". Thus much more needs to be said about external reference. But note that the causal theory does not in fact say much about eref; what is supposed to give terms their (external) meaning—dubbings—have not been explicated in any detail. Devitt (1981) has the most thorough discussion, but even there tends to assume the external object R that is dubbed is fairly obvious (e.g., physically present and attended to by P). And of course "attending to" is a largely internal matter. In fact our presentation of iref provides a portion of an account of attending: S points to (is linked to) R_P. Thus invoking of the S token might play the role of (immediately) attending to R_P (and mediately—and contingently—to R). Externally, it is totally indeterminate which of the many objects (or object parts) is attended or pointed to by S—the qua problem (Devitt and Sterelny 1987, pp. 63ff) arises precisely here. But the individual P doing the referring knows what she means, namely she knows she means R_P.

We can suggest a Generalized Kripke Condition: the user of a symbol *as symbol* must intend that it refer to something or other, i.e., that it be a stand-in. This is virtually a tautology, merely a definition of "symbol". But it is also instructive in that it focuses us in an internal direction, toward the meaning agent's own mental structures and how these serve to use one thing as stand-in for another. But this in turn says that there must be in the cognitive architecture not only tokens but also links between tokens, providing that something-or-other that a symbol S is to mean: something in P's notional world that makes S into a symbol.

To repeat: when we use a term semantically, we know its meaning by inner ostension; its meaning is a something-or-other; it means *that thing* we point to in our mental space. Not to ascribe meaning at all to a term is not to use it as an object with a semantics; and if we do not so use it, then it has no semantics, for we "meaners" are the creators of semantics. The mental pointing to an intended something-or-other is reminiscent of attention, and I think this is no accident. We think by attending, and this involves a directionality in which we direct our attention toward "that". This ties in with our remarks two paragraphs earlier about attending. Of course, this is merely suggestive, hardly a theory of attention.

"Arbitrariness", say Devitt and Sterelny (1987, p. 5), is one of the key features of language: "ln general, linguistic symbols have no intrinsic or necessary connection with their referents." Compare this to the opening quote from Fodor, where a similar thought is expressed. Symbols do not dictate their own meanings; we dictate them. I think that this feature of language, that *we decide what our words are to mean* rather than that the words carry their meanings with them, is the single most characterizing feature of language or symbolic behavior. We decide to link S to R_P and not to T_P, for instance. In stating much earlier that P might rename her dog, we alluded to this very phenomenon. And in deciding "Purdue" refers to a person (and not a university) Sue also invokes this power. Of course, we don't in each and every utterance re-decide meanings; we simply carry over most of them from our own past usages of the words in question.

To avoid misconstrual of my point, let me add that I do not intend to be saying here that we make a conscious decision about each word, even when we first learn it. And we might very well (a la Kripke-Putnam) be simply going along with what others seem to be doing with that word. But nevertheless we do form an internal link, and it is that link that determines what the word means to us, it is not the causal history per se that determines this. In other words, it is the Generalized Kripke Condition that is at work. This is also the aforementioned Gricean view accepted by and alluded to in (Devitt and Sterelny, 1987, pp. 119–127). Of course, in order for there to be effective inter-personal communication, these inner links must share certain features among persons in a linguistic community. What these features are, is hard to say, for ultimately this seems to depend on a theory of external reference, which I attempt in the sequel paper.

7 Error and Input Verificationism

If Sue thinks Purdue is that man over there, and she says “There’s Purdue” (pointing toward the returning figure and away from campus), is she right, or wrong? Well, she is wrong etymologically (i.e., her utterance is false with respect to historical accounts of meaning), but she is (largely) right conceptually. It’s just that she is naming things in a way different from others. Of course, in our story she does not realize this, and so is mistaken in her belief that she has correctly understood what others mean. But, if an argument were to ensue as to where Purdue is, in an important sense it would be an empty argument, for different things would be meant, and many of the underlying facts would not be in dispute. Yes, that man is over there; yes, campus is over here.

Now, this has ties to verificationism and hence to the problem of error. If what a person means by an expression is given by information in that person’s head, how can she ever be wrong about what she thinks? If what she means by “Purdue” is a person, how can she be wrong about Purdue (as she uses that term) being a person? Yet in a sense she certainly is wrong, and would probably admit to it once she learns of the public use of the term.

I suggest an “input verificationist” approach to this. It is a causal-descriptive approach, in which the descriptive component plays a much larger role than for Devitt and Sterelny. I argue that both reference borrowing and reference grounding have large doses of description to them. The key descriptive element is that of iref: the user associates the term in question with a meaning (for her use of the term). The key is that it allows predictions of behavior, as we saw earlier for Sue.

Now we return to the conflationary aspect: Sue describes Purdue to herself as *both* a person *and* as whatever the borrowees were talking about. She conflates two things. This allows her then to find out about her conflation and decide that her usage is inappropriate or confused and hence worth altering. She now decides to use “Purdue” the way the rest of us do, and regards her earlier definition as a poor one. That is, additional information (input) can convince her to adopt a new iref link, i.e., to change her internal meanings into ones that (again, in *her* notional world) fit the (new) facts better.

Note that had Sue not obeyed Kripke’s Condition, she would have no error to find out about, for she would have no reason to expect her usage to be like ours. And had she not obeyed the Generalized Kripke Condition, she would have no symbolic usage at all and so could not, for instance, later change her mind as to what she meant (there was nothing she meant), nor tell another what she takes Purdue to be.

There is still a matter to be explained: how does Sue ever relate her meanings to anything outside her, e.g., to other people’s usages, or to a real man with a cane? One answer is that perhaps she does not; she may simply wait to see what her experiences bring into her world view. Then if she finds that she has an unexpected experience, she can say she was wrong in her prediction. But this will not satisfy *our* demand for a semantics, a way for *us* to say that Sue’s beliefs are true or false. I will leave this here, since external reference will be treated in the sequel paper.

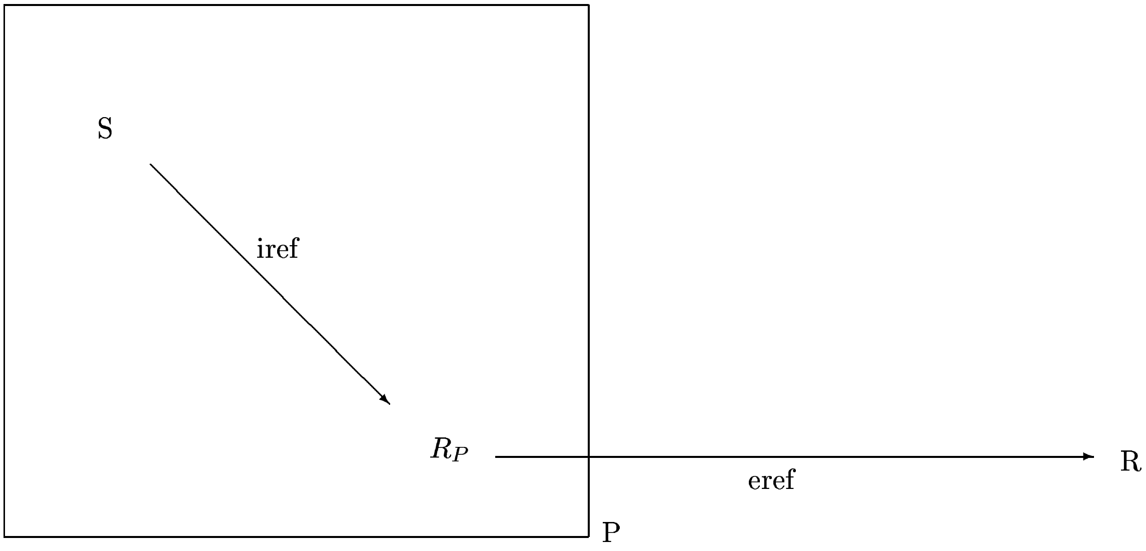

Figure 3: The Two-Tiered Model with Iref and Eref.

8 Conclusions

The picture is now two-tiered; there is additional structure in P's head, and of a special sort. An entire reference relation (although not a fully satisfactory one) is spelled out there. P has both symbol S and its referent (to her) R_P. R_P is typically produced (at least in simple cases) by an external entity R. The picture suggested now is as in Figure 3.

Here P is a genuine meaner: she *uses* S as a symbol for an object *known to her in her world view* as R_P. R_P may in turn be linked in a suitably reliable way to an external object R; giving an account of this remains for the sequel paper. But note that the causal theory does not provide such an account, contrary to claims for it. For the causal theory places the dubbing act in the sphere of internal acts (e.g., descriptions—Kripke is very explicit on this; more vaguely so are Devitt and Sterelny, p. 125). Just how it is that a name gets attached to an *external* object, even in the very act of ostensive dubbing, needs to be further explicated.

Iref then picks out a referent in our internal "mental space", which is all that a meaner can do. We are stuck with the conceptual abilities of our mind/brains. But for this to be *useful* there must be a suitably engineered connection between our inner world view (in our mental space) and the outer world. This latter connection, however, is not a linguistic one. There are no words involved!

However, there is an obvious question to be asked. Suppose that, as I have argued, we need internal referents in order to account for semantics-talk. But what good does it do? How is a two-tiered system better off than a one-tiered system? The former does allow us to account for error via input verificationism. But aside from the fact that we seem to find error in our thoughts and thus want to account for it, what does a system gain by being able to be in error? The ability to be in error goes hand in hand with the ability to be right, and both occur only in two-tiered systems. Furthermore a great advantage occurs to such systems: they can reflect on and change their views of the world. More on this below.

Moreover, iref is essential for the account of eref I will give in the sequel paper. Roughly, to get content out there, we first need to get it in here. To think about a foot, we need a foot-ish thing in our thoughts. This will not be simply the word "foot", for it is to supply our understanding of what a foot is, i.e., it supplies our meaning of "foot". And as stated above, it is flexible, in that we can decide to change the meaning when we see that it is less useful than some other meaning. Over time, we get new data, that can contradict our beliefs, for the latter often involve beliefs about future data. When future data suggest a better interpretation of past data, we can make adjustments and thereby avoid a strict verificationist dilemma.

We have mechanisms for trying to sort out things when confusion (non-veridicality) sets in. Thus "though [MacBeth] and I both make mistakes, we are both in a position to recover," as Fodor says (1990, p. 107). Recovery is a key notion, but Fodor does not dwell on it. Recovery is an inner comparison between what is and what isn't, in the agent's world. That is, between what the agent takes to be really out there, and what she takes to be in her mind. To do this she needs to have both of these represented in her mind, i.e., she needs a robust world-model that includes a self-model. The self-model is where the not-really-out-there things go: they are one's thoughts, feelings, imaginings, etc.

We can now state what a two-tiered system can do that a one-tiered system cannot. And in the process, we see the deficits inherent in a meta-linguistic incompetent (one who knows nothing *about* his language). Recall the claim by Devitt (1981) that a language-user need not have meta-linguistic competence; he need not know what words are, for instance, in order to use words correctly. But at what cost! Such an impoverished individual would be practically incapable of communication, indeed largely incapable of thought! (Here I will take the liberty of also assuming incompetence with respect to meta-ideas in general, i.e., the unfortunate being is devoid of the notion of any internal mechanisms in his behavior, and thus does not distinguish his thoughts as thoughts apart from what the thoughts are about. That is, instead of internally linking S and R_P, there is only S in the head, with of course R outside.) For consider the following behaviors, apparently unavailable to a one-tiered system but readily available to a two-tiered system:

1. Distinguishing a wanted possibility from the real thing.

2. Satisfying Kripke's Condition, upon hearing a new word; i.e., taking the word to mean whatever the others meant.

3. Asking or answering "What is your name?"

4. Discussing two persons named "John"; or one person with two names.

5. Asking or answering "Why did you say that?"

6. Asking or answering "What do you mean by X?"

7. Asking or answering "Don't you mean 'big' instead of 'small'?"

8. Recognizing that someone lied.

9. Recognizing mistaken ideas and trying to correct them.

10. Recognizing one's own ignorance and thus the need to find out new information.

I will not argue for each of these in detail. I hope the preceding discussion at least makes plausible that they are not available to a one-tiered system, but that they are to a two-tiered system. Another seeming benefit of this sort of consideration is that it suggests an intrinsic distinction between, say bee-language and human language. Devitt and Sterelny (1987, p. 5) can be construed as arguing that bees lack the "arbitrariness" feature (as well as the "stimulus independence" feature). I agree; however, the current approach puts arbitrariness (and therefore stimulus independence) in a central light whereby human language gains its great power. It allows us to think, to discover and correct errors; indeed, if I am right here, it is what gives our language meaning in the first place.

We need symbols in order to present possibilities to ourselves. That is, we need to be able to represent things and then on occasion distinguish the representations from the things represented, when we suspect that there may be a significant difference between the two. This is what allows us to represent uncertainty and error, and it amounts to recognition of inner processes largely but imperfectly attuned to outer events. In effect it amounts to recognition of ourselves as thinking beings.

Part II will attempt to provide an account of external reference along the following lines: An external object Q is related to the meaner (reasoning agent) S in much the same way that the notional object Q_S is related to S's self-notion S_S. Thus we introduce a self-notion in part for this purpose. The basic (easiest) case is that in which Q is part of S's body (e.g., Q is S's foot). Then the "in much the same way that" allusion two sentences above can be given physiological force via the actual pathways between foot and brain (e.g., leading to the tectum). Once we have a semblance of body-meaning along these lines, we can try to extend it further outward by a kind of "body geometry". (In this we are not far from the spirit of (Sloman 1986), (Steels 1986), and (Johnson 1987); see also (Perlis and Hall 1986), and (Perlis 1986, 1987, and forthcoming-a)).

9 Acknowledgments

Thanks to Michael Devitt, Georges Rey, Stephen Tighe, Alice Kyburg, Michael Miller, Rosalie Hall, Dorit Bar-On, Bill Lycan, Rich Thomason, Jeff Horty, Rolf Eberle, Brad Miller, Bill Hodos, Jim Reggia, John McCarthy, Dan Dennett, Pat Hayes, Robert Audi, Wolfgang Bibel, Bill Rapaport, Stu Shapiro, Andy Haas, Kathleen Akins, John Batali, Barbara von Eckardt, and all the others who encouraged, disagreed, listened or otherwise helped me in struggling with these ideas.

This research has been supported in part by the U. S. Army Research Office (DAAL03-88-K0087).

10 Bibliography

- Brentano, F. (1874) Psychologie vom empirischen Standpunkte. Leipzig.
- Dennett, D. (1987) The Intentional Stance. MIT Press.
- Dennett, D. (to appear) Granny's campaign for safe science. In G. Rey and B. Loewer (eds.) Meaning in Mind: Fodor and his Critics. MIT Press.

- Devitt, M. (1981) Designation. Columbia University Press.

- Devitt, M. (1989) A narrow representational theory of mind. In Representation: Readings in the Philosophy of Mental Representation; S. Silvers, editor. Kluwer.

- Devitt, M. (1990) Naturalistic representation. Manuscript.

- Devitt, M. and Sterelny, K. (1987) Language and Reality. MIT Press.

- Fauconnier, G. (1985) Mental Spaces. MIT Press.

- Fodor, J. (1979) The Language of Thought. Harvard University Press.

- Fodor, J. (1987) Psychosemantics. MIT Press.

- Fodor, J. (1990) A Theory of Content. MIT Press.

- Harman, G. (1975) Language, thought, and communication. In Gunderson, K. (1975) Editor. Language, Mind, and Knowledge. Univ. of Minnesota.

- Johnson, M. (1987) The Body in the Mind. University of Chicago Press.

- Kripke, S. (1980) Naming and Necessity. Harvard University Press.

- Locke, J. (1689) Essay concerning human understanding.

- Magee, B. (1988) The Great Philosophers. Oxford.

- Minsky, M. (1968) Matter, Mind, and Models. In: Semantic Information Processing, ed. M. Minsky, pp. 425-432. Cambridge: MIT Press.

- Perlis, D (1986) What is and what isn't. Symposium on intentionality, Society for Philosophy and Psychology, Johns Hopkins University.

- Perlis, D. (1987) How can a program mean? Proceedings, International Joint Conference on Artificial Intelligence, August, 1987, Milan, Italy.

- Perlis, D. (forthcoming-a) Intentionality and defaults. Advances in Human and Machine Cognition, K. Ford and P. Hayes (eds.). JAI Press.

- Perlis, D. (forthcoming-b) Putting one's foot in one's head—Part II: How. In J. Harper and A. Ramsay (eds.) Propositions and Attitudes: Philosophical Logic and Artificial Intelligence. Kluwer Studies in Linguistics and Philosophy.

- Perlis, D. and Hall, R. (1986) Intentionality as internality. Behavioral and Brain Sciences, 9(1), pp. 151-152.

- Putnam, H. (1975) Mind, Language and Reality: Philosophical Papers, vol. 2. Cambridge Univ. Press.

- Putnam, H. (1981) Reason. truth and history. Cambridge U. Press.

- Putnam, H. (1987) Representation and reality. MIT Press.

- Quine, W. (1961) From a Logical Point of View. 2nd edition. Harvard Univ. Press.

- Rapaport, W. (1976) Intentionality and the structure of existence. Ph.D. thesis, Indiana University.

- Rapaport, W. (1988) Syntactic semantics: foundations of computational natural-language understanding. In J. Fetzer (ed.) Aspects of Artificial Intelligence, Kluwer, pp. 81-131.