META-EXTERNALISM VS META-INTERNALISM IN THE STUDY OF REFERENCE

Daniel Cohnitz, Jussi Haukioja

Abstract

We distinguish and discuss two different accounts of the subject matter of theories of reference, meta-externalism and meta-internalism. We argue that a form of the meta-internalist view, "moderate meta-internalism", is the most plausible account of the subject matter of theories of reference. In the second part of the paper we explain how this account also helps to answer the questions of what kind of concept *reference* is, and what role intuitions have in the study of the reference relation.

Keywords:
reference, intuition, externalism, internalism, meta-externalism, meta-internalism, philosophical methodology

1. Introduction

In *Deconstructing the Mind*, Stephen Stich arrived at the following diagnosis of the philosophy of language as a discipline:

> [I]n most parts of physics or biology or archeology it is pretty clear what the theory is expected to do; though there may be a bit of squabbling about it from time to time, there is typically considerable agreement about the sorts of facts that a theory is expected to describe or explain. By contrast, it is far from clear what sorts of facts the theory of reference is supposed to account for. Indeed, it is my suspicion that, while the issue is only rarely a topic on which they have explicitly formulated views, different writers have quite different expectations. And, no doubt, some of the disagreement about which theories of reference are most promising can be traced to this underlying, largely tacit, disagreement about the job that a theory of reference is expected to do.
>
> [Stich 1996: 38]

We believe that Stich's diagnosis is still largely correct; there seems to be considerable confusion about the subject matter of theories of reference, which, it seems to us, also leads to further confusion about the proper methodology that semanticists should adopt in order to study reference.

In this paper we will distinguish different accounts of the subject matter of theories of reference that all seem to be prevalent in philosophical semantics (Sections 2-4). It will turn out that not all of these accounts of the subject matter of theories of reference mesh well with the explanatory role that these theories are supposed to play in a broader account of communication. In Section 5 we will elaborate an account that we call "moderate meta-internalism", which does seem to allow theories of reference to play an appropriate explanatory role. In Section 6 we will discuss the consequences this view has for the concept *reference* and for the role of intuitions in the study of reference. In the remainder of this introduction we will briefly review how disagreements about the subject matter of theories of reference have lead to disagreement and unclarity about the role of intuitions, as well as about the relevance of empirical results to theories of reference.

1.1. First-Order Internalism vs Externalism

Theories of reference are supposed to explain in virtue of what the referring expressions of natural language refer to the things that they in fact refer to. The descriptivist theory of reference is one such example (e.g. Searle [1958]). The theory maintains that, for example, a proper name in ordinary language refers to the object which uniquely or best satisfies the description (or a bundle of descriptions) that the person using the name associates with it. A well-known alternative to this view is the causal-historical theory of reference, often attributed to Saul Kripke [1980]. This theory states that the reference of at least some of the terms a speaker uses is (at least partly) determined by her physical and/or social environment. A proper name is introduced into the language in an act of baptism, which fixes the extension of the term. Reference is preserved in a causal chain of linguistic usage leading back to the original referent of the term, regardless of whether the speaker is actually aware of any part of this causal history.

A descriptivist theory is usually considered to be “internalist”[1] in the sense that the theory specifies internal, mental states of the speaker which determine the referent of the speaker’s utterance. The causal-historical theory, on the other hand, specifies circumstances outside the mind of the speaker that determine the referent of her expressions, and is thus considered to be “externalist”. In addition to Kripkean causal chains, familiar versions of externalism appeal to other external features such as the underlying natures of natural kinds [Putnam 1975] and the superior competence of experts [Burge 1979]. In our terminology this is a difference between *first-order externalism* and *first-order internalism* about reference:

> *First-Order Internalism*: The reference of any linguistic expression used by a speaker S is determined by the individual psychological states of S.
>
> *First-Order Externalism*: The reference of (at least some) linguistic expressions used by a speaker S, is determined (at least partly) by factors independent of the individual psychological states of S.

1.2 The Apparent Methodology of Philosophical Semantics

If confronted with two such fundamentally different theoretical proposals, one might begin to wonder which one of them is correct. Philosophers of language seem to have settled to a considerable extent on the second, externalist view. What seems to have convinced them to believe that externalism is the true theory of reference are *thought experiments*: hypothetical cases described to elicit intuitive judgments. In *Naming and Necessity*, Kripke presents a wide range of such hypothetical cases [Kripke 1980]. One of his most well-known examples concerns a counterfactual situation in which Kurt Gödel did not actually prove the incompleteness of arithmetic himself, but rather stole it from a colleague named “Schmidt”. However, according to the hypothetical story, most people who use the name ‘Gödel’ believe only one thing about him: namely that he is the person who proved the incompleteness of arithmetic. Would they refer to Schmidt when they use the name ‘Gödel’, as the descriptivist theory seems to predict? To Kripke it seems that they would not, and apparently the majority of philosophers of language share this reaction.

A few years ago, Edouard Machery, Ron Mallon, Shaun Nichols and Stephen Stich began to challenge this methodology [Machery et al. 2004]. They presented versions of the aforementioned thought experiment to ordinary speakers, and found that the folk’s intuitions

[1] In this paper we will use double quotation marks when using a quoted expression, and single quotation marks when mentioning an expression.

vary considerably about this case. On this basis, Machery et al. call for a revision of philosophical practice. Intuitions about thought experiments should no longer be considered evidence for theories of reference, if such intuitions seem to vary unpredictably within the linguistic community. Since there is (apparently) no other evidence for theories of reference, Machery et al. suggest that theorizing about reference should cease and a form of deflationism about reference should be adopted [Mallon et al. 2009].

This argument has been at the centre of discussion for some time now. Philosophers of language have pointed out flaws in the experimental setup of the initial experiment, some of which were fixed by later experiments. Others pointed out flaws in the overall argument by Machery et al., some of which were addressed by the experimentalists. However, one of the peculiar aspects of the critical discussion of Machery et al.'s study is that there does not seem to be a consensus within philosophical semantics whether the semantic intuitions of ordinary speakers are relevant data *at all*.

Thus, some philosophers of language have argued that even if that variation existed, it would not be problematic for theorizing about reference in the traditional way, because the variation was found in intuitions of lay-persons, and their intuitions are not very relevant. The intuitions that count in determining what is true about reference should rather be the intuitions of experts, as is the case in other domains of science:

> What then should we make of referential intuitions? And whose intuitions should we most trust? . . . Still, are [the intuitions of ordinary competent speakers] likely to be right? I think we need to be cautious in accepting them: semantics is notoriously hard and the folk are a long way from being experts. Still it does seem to me that their intuitions about "simple" situations are likely to be right. This having been said, we should prefer the intuitions of semanticists, usually philosophers, because they are much more expert (which is not to say, very expert!). Just as the intuitions of paleontologists, physicists, and psychologists in their respective domains are likely to be better than those of the folk, so too the intuitions of the semanticists.
>
> [Devitt 2011a: 425-6]

Other philosophers of language, though likewise very sceptical about the claims made by Machery et al., apparently believe that the intuitions of ordinary speakers matter, but that the specific intuitions tested by Machery et al. happened to be the wrong ones:

> [Machery et al.] test people's intuitions about *theories* of reference, not about the *use* of names. But what we think the correct theory of reference determination is, and how we use names to talk about things are two very different issues.
> In testing people's intuitions, I think it is important to distinguish carefully between observations that will reveal how people do things (in this case, use names) and observations designed to reveal how they think they do them. The latter will only provide grounds to determine how they are disposed to theorize about their practices, i.e. predict which theories about what they do they are disposed to favour.
>
> [Martí 2009: 44]

To some extent, the apparent disagreement in the verdicts by Devitt and Martí in those quotes can be traced back to their different use of the term 'intuition'. Devitt is only using it in the sense of meta-linguistic intuition that reveals what people *think* about their usage, whereas Martí's use of the term also covers intuitions that reveal how people actually *do* use and interpret expressions.[2] Nonetheless, as we shall show below, Devitt's background assumptions regarding the determination of reference lead him to the view that intuitions, even in the wider sense, have no constitutive role to play in reference:

> It is common to think that the task of, for example, the theory of reference is simply to systematize our ordinary intuitions about reference. . . . Still this common view is puzzling. It is puzzling because the obvious way to describe the task of the theory of reference is to explain the nature of reference, to explain the nature of a certain word-world relation. If we start from this view, surely as good a starting place as one could have, why take the task to be to capture the folk theory of this relation? . . . Why think think that the folk have particular insight into the nature of this particular word-world relation? We don't suppose that they are authorities on physics, biology, or economics, why suppose that they are on semantics?
>
> [Devitt 2012: ??]

To make progress on this issue, we need a better understanding of what, exactly, theories of reference are theories of, and what determines the correct theory of reference. Only after these matters are settled does it make sense to ask which evidence is the most appropriate for philosophers of language to use as a basis for their theory choice.

We suggest that it is helpful, in clarifying these issues, to draw another internalism–externalism distinction in addition to the familiar one described above. The distinction we have in mind (between what we will below call "*meta*-internalism" and "*meta*-externalism") is concerned with what determines the correct theory of reference: is it determined by factors internal to individual speakers, or is it—as in Devitt's quote—just as mind-independent as biology or physics?

2. Häggqvist and Wikforss on "A Posteriori Semantics"

A distinction that is in many ways similar to the one we are making has been suggested by Sören Häggqvist and Åsa Wikforss [2007]. Häggqvist and Wikforss claim that externalist thinking in semantics, and the idea that meaning can partly depend on features of the external environment to which we only have a posteriori access, has been extended too far by some theorists.

> [A] more radical thesis has emerged, a thesis we shall dub 'a posteriori semantics'. The suggestion is that not only does the meaning of a term *t* depend on the external environment, in classic Putnam-Burge fashion, but also that the semantics of *t* depends on the external environment. For instance, it has been argued that whether or not 'water' should be given an externalist semantics or a descriptivist one, depends on facts about the physical environment, such as facts about chemical composition and microstructure.
>
> [Häggqvist and Wikforss 2007: 375]

[2] Devitt uses 'intuition' to refer to "fairly immediate unreflective" but nevertheless theory-laden judgments about metalinguistic questions [Devitt 2006: 95]. We suggest distinguishing between the following: (i) the intuitive interpretation of expressions in utterances of actual or hypothetical communication partners and the intuitive selection of expressions for use in actual or hypothetical communication situations (this is what Devitt seems to call "performance"), (ii) a subject's reports about her intuitions in the sense of (i), (iii) reflected interpretations and selections, (iv) a subject's reports about (iii). Devitt seems to complain that (iv) is problematic evidence for theories of reference if it comes to folk judgments, and that a theory of reference should not be considered to be a systematization of such intuitions, and he is probably right about this. We suggest in this paper, however, that the methodological issue is about the reliability of (ii) as evidence about (i), and that (i) (insofar as it is an output of the subject's linguistic competence) should be considered constitutive for reference. This seems to us nevertheless to be in disagreement with Devitt's view.

Häggqvist and Wikforss argue against A Posteriori Semantics based on two kinds of considerations, one metaphysical, the other epistemological. The metaphysical complaint is that A Posteriori Semantics makes facts about meaning completely dependent on facts that are not merely external to the speaker but completely remote from our linguistic practice. But how could those facts possibly play this meaning-determining role?

> The thesis [of A Posteriori Semantics] is that microphysical *facts* determine the semantics of common, vernacular terms used for centuries . . . This alleged "metaphysical" dependence is *the reason* why it is a posteriori what sort of semantics a sentence (or thought) has, according to the thesis' proponents. And it is simply obscure how the physical details of the world could play such a pivotal role. Facts about chemistry, it would seem, do not carry that kind of semantic significance—all by themselves.
>
> [Häggqvist and Wikforss 2007: 380]

The first of their epistemological complaints is closely related to this metaphysical complaint. Since it is possibly a matter of remote microphysical details that determines the semantics of our expressions, knowledge of these remote details is required for knowledge of semantics. A subject living prior to the eighteenth century, ignorant of modern chemistry and modern physics, would simply not be in the position to know the semantics of (some of) her expressions, regardless of whatever other linguistic knowledge she might have.

The second epistemological point is methodological in character. Häggqvist and Wikforss note that externalism is typically argued for with the help of thought experiments, eliciting modal intuitions that speak in favour of one semantic theory rather than another. But how can speaker intuitions be of any relevance for determining which semantic theory is correct, if the latter is completely determined by facts external to the speakers and possibly unknown? It seems that A Posteriori Semantics undermines the very methodology it is based on.[3]

We agree with Häggqvist and Wikforss that these metaphysical and epistemological considerations are powerful objections against certain views concerning the determination of reference, but we do not believe that construing the problematic position as "A Posteriori Semantics" is the best way to make that point. The distinction between a problematic A Posteriori Semantics and (a presumably less problematic) *a priori* semantics is unfortunate, perhaps even misleading, in two respects:

Firstly, the whole issue does not seem to be about whether it is *a priori* or *a posteriori* which theory of reference is true, because also the "*a priori* externalist" as well as the "*a posteriori* internalist" would have to consider what a term refers to in the mouth of another speaker as a two-fold *a posteriori* matter, since (a) it depends on the intentions of the other speaker how she intends to use the term, and (b) it then depends on further facts that determine what the term refers to, depending on what term it is. Only the types of the expressions in ones own utterances are under ideal conditions *a priori* knowable to oneself, if there is such a thing as *a priori* knowability. Moreover, the notion of '*a priori*' that Häggqvist and Wikforss use in their characterization of A Posteriori Semantics is—as Häggqvist and

[3] Häggqvist and Wikforss raise a third epistemological complaint, viz. that A Posteriori Semantics also makes logical form and hence validity and other inferential relations between expressions accessible only *a posteriori*, which would render our reasoning abilities in an implausible way depending on empirical science. Since this complaint is specifically directed at one theory (Ludlow's externalism about logical form, discussed below), we will leave this out of our discussion and concentrate on the objections raised against A Posteriori Semantics in general.

Wikforss note themselves—somewhat problematical, since it counts all knowledge about a speaker's inner (non-environmental) states that are accessible through introspection, and all knowledge justified on the basis of memory to be "*a priori*".

The second way in which the distinction is unfortunate is that it counts as "*a posteriori*" also those theories that suggest a dispositionalist but conditional analysis of meaning. Take a theory that would state that the speaker's dispositions to apply 'water' determine that the term has the semantics of a natural kind term and refers to the underlying chemical substance if there is one, but adopts the semantic profile of a functional kind term, if there is no such unique structure. Let's further assume that this is not merely the way ordinary speakers intend to use the term, but that this is actually *a priori* knowable about the term 'water' (for accounts along these lines, see Korman [2006] and Haukioja [2009]). Still, this theory is considered an instance of A Posteriori Semantics by Häggqvist and Wikforss. However, the objections formulated above against A Posteriori Semantics miss their target, if A Posteriori Semantics is supposed to also include such a position. First of all, the semantics of 'water' would now be determined (at least in part) by speaker dispositions, not by microphysical environmental facts alone. Thus the metaphysical objection loses its bite. Similarly, the conditional semantic facts are *a priori* knowable, and thus the semantics become scrutable in a plausible way (true, it is still not *a priori* knowable that 'water' refers to H_2O, or that 'water' refers to a unique kind, but this is a feature, not a bug of the account).[4] So the first epistemological objection loses its force. The second epistemological objection, according to which A Posteriori Semantics undermines the very methodology it is based on, also does not apply to the conditional dispositionalist. Since it is speaker dispositions to apply terms in actual and counterfactual scenarios that determine the conditional semantics, it is plausible that responses to thought experiments shed light on the conditional semantic profile of our terms.

3. Meta-Externalism vs Meta-Internalism

We believe that taking our earlier distinction between first-order externalism and internalism to the meta-level captures the intentions of Häggqvist and Wikforss more effectively. Meta-internalists hold the view that the conflict between first-order internalism and first-order externalism should be resolved by appealing to the psychological states of speakers (the intuitions, mental dispositions or representations, etc., of the speaker determine which theory of reference is correct for her linguistic expressions), whereas meta-externalists hold that even this can be determined externally.

> *Meta-Internalism*: How a linguistic expression E in an utterance U by a speaker S refers[5] and which theory of reference is true of E is determined by individual psychological states of S at the time of U.

[4] An anonymous referee has objected that conditional accounts will still make it an external matter—and thus implausibly hard to know—whether a given term has an externalist or internalist semantics. But we think this objection mischaracterizes the conditional view. On such a view, the fact that a given term—say, 'water'—has externalist semantics will be determined by individual psychology: it is just that under certain kinds of external conditions 'water' will refer, not on the basis of microstructure, but on the basis of some functional or manifest properties. It will not, then, be *a priori* knowable whether 'water' refers to H_2O, to a finite disjunction of underlying kinds, or to anything satisfying a certain functional description—but this is not any more objectionable on epistemological grounds than the general externalist claim that it is not *a priori* knowable whether 'water' refers to H_2O or XYZ.

[5] 'How a linguistic expression E refers' is our shorthand for, roughly, 'whether external factors can play a role in determining what E refers to, and if so, which kinds of external factors: causal chains, or underlying essences, or the judgements of experts, or . . .'

> *Meta-Externalism*: How a linguistic expression E in an utterance U by a speaker S refers and which theory of reference is true of E is not determined by the individual psychological states of S at the time of U.

Although meta-externalist views have been put forward by theorists who also subscribe to first-order externalism, it is important to note that the two internalism–externalism distinctions are logically independent. The first-level distinction concerns the question of what kinds of properties enter into the determination of what a linguistic expression refers to. The meta-level distinction concerns the question of *why* these properties play this role—what makes it the case that a particular linguistic expression refers in the way that it does. Even if internalist and externalist views (or prejudices) on both levels perhaps tend to go hand in hand, there does not appear to be any immediate incoherence in combining meta-internalism with first-order externalism, or meta-externalism with first-order internalism. In fact, we believe there are good grounds for accepting the first combination, of meta-internalism and first-order externalism. *What* a term refers to is in many cases determined by external factors, but *how* it refers—whether external factors in fact *do* play a role in determining its reference—is determined by the individual internal states of the speaker.

In the next section we will argue that meta-externalism is an implausible view about semantics. We will then, in the fourth part of this paper, turn to a discussion and elaboration of meta-internalism.

3.1 Meta-Externalism

It seems to us that meta-externalist views are often taken to simply follow from first-order externalism. The familiar thought experiments that show that the referents of at least proper names and natural kind terms are externally determined, are also—without further argument—assumed to show that *how* terms refer is also externally determined. We believe that this is mistaken, and that carefully distinguishing between the first-order distinction and the meta-level distinction is crucial in helping to see and avoid the mistake.

Because these two levels are typically not distinguished, it is not always clear whether externalist ideas are taken to be restricted to what we have called first-order externalism, and when the further commitment to meta-externalism is being made. Accordingly, explicit statements of meta-externalism are difficult to find in the literature. But we can point to some cases where a commitment to meta-externalism is clear. A radical case of meta-externalism will be discussed below, in 6.2, when we turn to Cappelen and Winblad's argument that the meaning of 'reference' is externalistically determined, which proceeds by exaggerating the conclusions to be drawn from externalist thought experiments.

Peter Ludlow's "bald externalism about logical form" is another obvious case of meta-externalism [2003, 2011]. Ludlow is concerned with the externalist's difficulty in dealing with empty names, as in 'Santa Claus delivered the toys'. On a descriptivist theory it is possible to say that such sentences are meaningful (but false) even though 'Santa Claus' does not refer. But what should an externalist say about such cases? On standard externalist accounts, one seems forced to say that 'Santa Claus delivered the toys' does not express a determinate proposition. Ludlow suggests "bald externalism" as one possible way of dealing with this problem: some names are referring expressions (as the externalist standard theory has it), while other names are descriptive (as the descriptivist theory has it). Which semantics a name has is completely determined by external factors:

> We make certain utterances and the logical forms of those utterances are what they are as fixed by the external world, completely independently of facts about our linguistic intentions.

[Ludlow 2003: 405]

Ludlow recognizes that cutting off the logical form of utterances from the psychological states of speakers is highly implausible. However, the fix that Ludlow offers is still meta-externalist. He simply gets the speakers' "intentions" back into the picture by individuating these intentions widely. Ludlow considers two molecular duplicates, Peter on Earth and Twin-Peter on Twin-Earth. Twin-Earth differs from Earth only in the fact that Socrates is an invention of Plato there. Now when Peter thinks 'Socrates was a philosopher', Peter's thought (being a singular proposition) differs in form and content from the thought that Twin-Peter expresses with 'Socrates was a philosopher' (being a general proposition), since they are living in different environments. According to externalism about logical form, also the linguistic expressions are different in logical form (one containing a referring expression, the other containing a denoting expression), but now correlated to the widely individuated thoughts of the speakers [Ludlow 2003: 407]. However, if semantics is determined by widely individuated psychological contents but not determined by the (narrow) individual psychological states of a speaker, then this is still an instance of meta-externalism as we have defined it.

Michael Devitt is another clear example. His views on intuitions, which we commented on above, already seem to hint at meta-externalism. The commitment is more explicitly visible in, for example, his discussion of deference (or "reference-borrowing", as he prefers to call it):

> If a person's current use of a name is to designate its bearer then that use must be caused by an ability with that name that is, as a matter of fact, grounded in the bearer whether via reference borrowing or directly by the person herself: the efficacious mental state must have the right sort of causal history. If it has the right history, that is sufficient.
>
> [Devitt 2011b: 202]

The right kind of causal history is sufficient for a name to refer, via a historical-causal chain, to its bearer. This does not need to be grounded in any *current* mental states of the speaker.[6] This makes Devitt's view, too, a clear instance of meta-externalism.

3.2 What is Wrong with Meta-Externalism?

The first problem with meta-externalist accounts arises when one considers the explanatory aims of (philosophical) accounts of semantics, one of which is, without any doubt, the nature and possibility of intersubjective communication. At least since Gottlob Frege, theoretical choices in semantics have been driven by considerations about how well the semantic theory is able to explain successful communication. Frege's own principle of compositionality, as a constraint on theories of meaning and reference, is motivated by the assumption that the phenomenon of successful communication with novel sentences could otherwise not be explained (cf. Pagin and Westerstahl [2010]). Of course, a theory of meaning and reference is not itself a theory of communication, but it is common in philosophy of language to assume that it systematically contributes to a theory of communication. As we will argue in this section, it seems that reference cannot be construed as being independent of speakers' individual psychology, as meta-externalists would have it, precisely because it plays this role

[6] In his discussion, Devitt [2011b: 202-203] seems to assume that the alternative view would require speakers to have explicit deferential ("backward-looking") *intentions*, or to acknowledge the causal history of the term. We agree with Devitt that this would be an over-intellectualized picture; however we do think that deference has to be grounded in current dispositional states of the speaker. A full discussion of this question is beyond the scope of the present paper, but we will briefly return to this in Section 5.

in communication. The explanation of the latter phenomenon cannot proceed in isolation from the psychological features of language-speakers.

In order to illustrate this, it may be useful to consider the following thought experiment. One might imagine a world where the correct theory of reference is determined by external facts, i.e. a world in which meta-externalism is true. One might further imagine that it is populated by speakers of Frenglish—they speak a language like English but have psychologically internalized a Fregean, descriptive theory of proper names. Thus, they consistently and intuitively use proper names to refer to the unique object, if there is one, that satisfies the description which the speaker associates with the proper name, and interpret each others' use of proper names accordingly. They also show the intuitive reactions to thought experiments (like Gödel/Schmidt-experiments) that the descriptive theory would predict. However, in this world the external facts about reference are such that a causal-historical theory of reference happens to be true. Since we are assuming meta-externalism to be true, this combination of facts would be coherent—the fact that the speakers intuitively *use and interpret* names according to the descriptive theory has nothing to do with which theory of reference is true.[7]

The problem is, of course, that such externally determined semantic facts seem to be completely irrelevant for the explanation of communication in Frenglish. The internalized descriptive theory of reference poses no obstacles for the communication amongst the Frenglish speakers: they successfully use names to communicate information about the individuals they *take* their names to refer to. But since externalism in fact is true in our hypothetical world, such communication succeeds through the use of sentences that are *systematically false* in a range of cases. In fact, in situations where the descriptive theory and the causal-historical theory give different verdicts, Frenglish speakers successfully communicate information using sentences which may not only be false, but which characterize individuals that have nothing to do with the information that gets communicated. Indeed, meta-externalism allows for the situation where all Frenglish speakers are always mistaken about the reference of their terms (even under otherwise idealized epistemic conditions).[8]

Assuming a separate linguistic reality creates additional methodological problems, namely: how does one go about studying it? How would one have access to it in the first place? If there are no constitutive relationships whatsoever between reference and our intuitions or our dispositions to apply and interpret terms, one can only wonder what should be taken as reliable evidence for referential relations. Since the traditional method of

[7] Some meta-externalists might object that the thought experiment assumes that everyone in the hypothetical world is a Frenglish-speaker, but that their version of meta-externalism would require that experts or some other relevant group *have* the relevant dispositions or intentions, and that it could only be because of this that (first-order) externalism is true in that world, even if not all speakers in that world need to have those dispositions. In order to accommodate this, one can assume that in our hypothetical world all "experts" died five minutes ago.

[8] One might argue in response to this thought experiment, and in the spirit of Ludlow's externalism about logical form discussed in 3.1, that the meta-externalist has further resources for dealing with the apparent miscommunication between Frenglish-speakers. If, in addition to being externalists about linguistic meaning in our hypothetical world, we also adopt externalism about mental content, it is not any longer that clear that the Frenglish-speakers do not successfully communicate with each other. At least, their mental states would now match the meaning of their utterances. However, this hypothetical world should still be a mysterious place for even such a sophisticated meta-externalist, since—although (on this account) Frenglish-speakers mean what they say and understand what was meant—they very often do not act accordingly at all. Hence, even if one assumes, with the sophisticated meta-externalist, also an externalist account of mental content for the hypothetical world, linguistic communication leads to mapping thought contents, but it still doesn't lead to communicative success in terms of coordinated behaviour, because the latter will, arguably, still depend on the narrow psychological content of our Frenglish community.

intuition-mining can only go as far as folk semantics [Devitt 2009] and we have no reason to assume that the “folk” are always right about reference, then the study of linguistic reality should proceed by methods alien both to linguists and philosophers of language. In addition, it would be hard to deny that we *do* know more about reference now than we did 50 years ago, before Kripke, Putnam, and others developed their (first-order) externalist theories. If the methods we have been using are as inferior as the meta-externalist picture would claim them to be, how was this even possible?

4. Meta-Internalism

If meta-externalism does not seem to be a promising view about reference, then perhaps we should consider its alternative. It is often suggested that the alternative to an externalist view about language is a view that locates all relevant meaning and reference determining facts deep in the structures of human mind/brains:

> According to internalist conceptions of language, languages are properties of the mind/brains of individuals and supervene entirely on the internal states of these mind/brains. Hence, languages are primarily to be studied by the mind and/or brain sciences—psychology, neuroscience, and the cognitive sciences more generally (including linguistics and philosophy).
>
> [Bezuidenhout 2006: 127]

This suggests a broadly Chomskian conception of internalism, according to which languages are not the social objects we naively consider them to be, but in fact properties of our individual brains. We will call this view “*radical* meta-internalism”.

4.1 Radical Meta-Internalism

> *Radical Meta-Internalism*: How a linguistic expression E in an utterance U by a speaker S refers and which theory of reference is true of E is determined by internal subconscious representations of the semantic theory in the mind/brain of S at the time of U.

Let’s consider two examples of radical meta-internalism. In his [2001], Gabriel Segal sets out to explain and assess two semantic theories of proper names within what he calls “a realist cognitivist framework”. He constructs the subject-matter of linguistics along Chomskian lines, adopting the position that linguistics is a part of psychology which studies linguistic competence arising from tacit representations of grammatical rules. Segal extends this model to semantic competence, which is the result of (again, largely unconscious) cognizance of a compositional semantic theory. Such internalized T-theories are supposed to be real natural phenomena. Thus, the meaning and reference of a speaker’s expressions are determined by her internalized, yet consciously unavailable representation of a semantic theory.

The two theories that Segal discusses can both deal with the classical issues concerning proper names (rigidity, co-extensionality and empty names) on an equal level of success, but differ in the computational mechanism by which they arrive at the semantic evaluation. In order to decide between these two theories, one would need to appeal to some deep underlying psychological facts.

A similarly radical meta-internalist approach is adopted in Stainton’s [2010] discussion of the bearing of psychological studies of proper names on descriptive theories of reference. The evidence from such studies seems to suggest that proper names are psychologically special—for example, they take longer to process and are more difficult to remember than descriptions. Since the psychological profiles of purportedly synonymous names and descriptions are different, the evidence from empirical psychology threatens the plausibility of descriptive accounts of reference. The major assumption behind this kind of approach is that semantic hypotheses can and should be tested by psychological means, because the reality, which these hypotheses set out to explain, is itself of a psychological nature. Even more so: a theory of reference is part of psychology proper since a true theory of reference and a true theory of meaning for a given speaker should be discoverable as a subconscious representation in the speaker's mind/brain.

4.2 What is Wrong with Radical Meta-Internalism?

Consider the "Martian argument" often invoked (e.g. in Devitt and Sterelny [1999]) as an objection to a Chomskian conception of language. Suppose a Martian succeeds in learning to speak English by internalizing a relevant set of linguistic (and semantic) rules. It seems natural to assume that the Martians' psychological constitution could be different from ours—perhaps they do not have a language faculty at all as a separate module. Now, on the Chomskian view, an alien with a different psychological make-up could not be speaking English, regardless of how well she seemingly manages this task and how successfully she can communicate to English-speaking Earthlings. This, however, is counterintuitive: it seems strange to deny the Martian's ability to speak English (or any human language, for that matter) simply because the creatures around where she lives happen to have a different psychology.

A similar argument can be used to cast doubt on radical meta-internalism. Suppose the Martians had (or developed) dispositions to use and interpret names that were exactly identical to our dispositions, but these dispositions were the result of very different psychological processes, due to differences in internal wiring between the Martians and us. Radical meta-internalism would now claim that quite different theories of reference were true of our usage of names, and of the Martians' usage of names. But this conclusion seems to just overlook the interesting *similarities* between us and the Martians; precisely those similarities which would enable us and the Martians to successfully use names to communicate information about individuals in the world. To be sure, the differences would be interesting and worth investigating, too. But as far as the explanation of communication by using names, and the determination of truth conditions for sentences including names are concerned, the differences seem to be irrelevant; requiring theories of reference to be sensitive to such details of internal wiring would be to miss the most interesting generalisations, those covering both us and the Martians.

Discussing thought-experiments involving aliens may seem like a contentious activity; one may try to dismiss the Martian argument as a far-fetched piece of science-fiction, that has nothing to do with the actual study of human languages [Laurence 2003]. However, a similar (and, actually, stronger) line of argument is available even for those who are sceptical about the linguistic abilities of extra-terrestrials.

Consider the acquisition of a second language. A consistent Chomskian would be forced to claim that each language-speaker possesses one internalized theory for her native language, and then another one for her second language. This would mean that every non-native speaker of English would end up being in the same position as the Martians, because English grammar and formal semantics would not be describing her linguistic behaviour due to a difference in underlying psychological representations. Whatever linguists and formal semanticists are saying about English would apply only to native speakers of English. Of course, from the perspective of psycholinguistics, distinguishing languages in this way might make perfect sense. But at the level of analysis at which we take the philosophy of language to be interested in meaning, reference and content, these differences do not seem to matter.

5. Moderate Meta-Internalism

If both meta-externalism and radical meta-internalism are implausible, we should look at the alternatives. We submit that the most sensible alternative is the view we call *moderate* meta-internalism:

> *Moderate Meta-Internalism*: How a linguistic expression E in an utterance U by a speaker S refers and which theory of reference is true of E is determined by S's dispositional states to apply and interpret E in actual and hypothetical circumstances.

According to moderate meta-internalism,[9] what makes it the case that a given expression is to be given an internalist or externalist semantics is our *patterns of application and interpretation*.[10] For some expressions, such as 'bachelor', our intuitions about proper application and interpretation are unaffected by contingent features of the actual world around us—the properties associated with the expression are taken to be decisive for the question of whether the expression applies to a given individual or not. With other expressions, if first-order externalism is correct, we have dispositions to "shift the burden" of determining their applicability partly to external factors. In the case of 'water', for example, we have dispositions to evaluate the correctness of actual and counterfactual applications of the term according to whether or not the term is applied to samples which share a microstructure with the substance that is causally connected in the appropriate way to our actual (past and present) usage of 'water'. (Or something similar: the precise details will of course depend on the correct formulation of first-order externalism for natural kind terms.)

The distinction between first-order internalism and externalism arises, then, in a very natural way within moderate meta-internalism: it is precisely *because* our dispositions to apply and interpret natural kind terms are different from our dispositions to apply and interpret terms like 'bachelor' that the former get an externalist semantics while the latter get an internalist one. It is important to note that meta-internalism does not require that speakers have explicit "burden-shifting" (or deferential; see below) *intentions*, let alone that the speakers semantically associate causal descriptions to externalistically referring terms. All that is required is that, when a term does refer externalistically, this must be grounded in the speakers' dispositions (for example, to revise usage in light of new information about the world).

We do not think or claim that meta-internalism is a highly *surprising* doctrine—something like it seems to be implicit in a lot of thinking and theorizing about reference and language. Since the assumption is implicit, it is typically not argued for—but the assumption is apparent in the typical methodology used in theorizing about reference. Typically, the question of whether a given expression is to be given an internalist or an externalist semantics is approached through thought-experimentation, where intuitions about the proper application and interpretation of the expression in question are elicited.[11] The choice of such methodology, especially if it is assumed to be a sufficient methodology for *completing* the project, appears sensible only if something like moderate meta-internalism is assumed. It is

[9] From here on, we will for the sake of brevity use the label 'meta-internalism' to refer to moderate meta-internalism, using 'radical meta-internalism' when specifically discussing that variety.

[10] It is worth noting that these patterns include our *systematic readiness to correct ourselves,* for example in the light of new information. Thus, not all dispositions are equal; any particular application can turn out to be mistaken. This qualification is also important in order to draw the necessary competence–performance distinction.

[11] As far as the assumption of moderate meta-internalism is concerned, it does not make a difference whether the intuitions are self-elicited in the armchair or elicited from other subjects in an experimental setting. We will briefly comment on experimental methods below.

worth noting that the two alternatives discussed earlier, meta-externalism and radical meta-internalism, agree that intuition-based methodology is at best a source of tentative evidence for facts about reference—both assume that there are other sources of data which can give us more direct or reliable access to such facts. According to moderate meta-internalism, this idea is misguided: reference is wholly determined by our dispositional states, so evidence about our dispositional states is not in any sense inferior to other kinds of data. Indeed, data about our dispositional states seem to be the best kind of evidence we *can* have. However, moderate meta-internalism alone does not tell us what kind of evidence is *good* evidence of our dispositional states, and whether dispositions should be studied using armchair methods or by experimental means. We will say more about this question in the next section.

We pointed out above how first-order physical externalism can arise on the meta-internalist picture. First-order social externalism arises in a similar fashion, through deferential dispositions. For example, I can refer to elms with my term 'elm', on the basis of my dispositions to defer to people who can actually tell elms apart from other trees (e.g. botanists or gardeners). In such familiar cases, it is reasonable to assume that, although I do not know enough about elms to refer to them "on my own", I do know that elms are trees, and that 'elm' is a natural kind term.

It might be objected that requiring such deferential dispositions of speakers is implausible, and that it is in conflict with our actual practices of attributing competence with deferentially referring terms such as proper names. Consider the following hypothetical case discussed by Louis deRosset [2011] as an objection to neo-descriptivist theories of reference: Ethan is a three-year old child who learns as part of his religious education in the kindergarten that Peter was an important man who told a lot of people about Jesus. DeRosset argues that Ethan thereby acquires sufficient facility with the name 'Peter' in order to refer to the apostle Peter. However, since Ethan is a three-year old kid, one might well doubt that he has the same meta-dispositions to adapt his usage and interpretation of 'Peter' to new information as normal adult ordinary speakers have:

> Deferring to others would require Ethan to associate with 'Peter' some metalinguistic or metacognitive condition, such as *the man the people who taught me 'Peter' intended to talk about using the name*. Empirical research suggests that a typical three-year-old like Ethan lacks the requisite metalinguistic and metacognitive abilities. . . . He cannot yet . . . pass the False Belief Test. Thus, he lacks a robust capacity to answer questions that depend on how another's view of things differs from his own. Importantly, he can't reliably differentiate a word from its referent.
>
> [deRosset 2011: 6]

Let us assume for the sake of exposition that in Ethan's usage 'Peter' does refer to the apostle Peter (if there is such) just as it does in our usage, since he acquired the name in the right way from competent speakers whose usage of 'Peter' does refer to the apostle Peter, even though Ethan is not (yet) a fully competent speaker (which might involve acquiring meta-dispositions that he is not capable of developing yet). A case like this, in which the reference of Ethan's usage of 'Peter' depends on the meta-dispositions of the fully competent adult speakers in his linguistic community, is *not* compatible with meta-internalism as we have defined it.

However, it is doubtful whether the above description of this case is adequate. From the perspective of meta-internalism, it seems more plausible to say that Ethan's usage of 'Peter' is different from the usage of fully competent speakers. It might be that 'Peter' in Ethan's idiolect happens to refer to the same individual that 'Peter' refers to in English. However, *how* the term refers in Ethan's idiolect and in English is different. It is implausible to think that if Ethan cannot even conceive that ‘Peter’ might refer to an individual he knows nothing about, his usage of ‘Peter’ nevertheless may refer via mere causal chains to that individual.[12]

6. Consequences

Meta-externalism and radical meta-internalism are both associated with particular views about the subject matter of theories of reference—that is, the nature of referential properties. That a particular token of a word refers to, say, Barack Obama, would be thought of as being an objective property of the token, not determined by the psychological states of the speaker uttering it (meta-externalism) or a property of the representational state of the speaker (in radical meta-internalism). However, it is not immediately clear what a moderate meta-internalist should take referential properties to be. In this final section we will take a look at this question, as well as the related question of what kind of methodologies are best suited for the study of reference.

6.1. The Subject Matter of Theories of Reference and the Concept Reference

Moderate meta-internalism claims that the truth concerning how a given term refers—including the question of whether its reference is determined internalistically or externalistically—is determined by the dispositional states of competent speakers (of the language that the term is part of).[13] This suggests that the referential property of a *term*—that it refers to something—should also be understood in dispositional terms.[14] For example, the property of referring to Barack Obama is, in a very rough approximation, the dispositional property of being such as to be *interpreted*, in favourable circumstances, as referring to Barack Obama by competent speakers in their linguistic activity.[15]

Theories of reference are not, of course, primarily concerned with the explanation of how a given *particular* expression refers. Rather, they are theories about how members of a *type* of expression refer. For example, our dispositions to apply and interpret proper names are remarkably uniform: all ordinary proper names at least appear to become connected to their referents by very similar mechanisms. At the same time, the reference of other types of expression may be determined by quite different mechanisms. It is not a coincidence that theories of reference are invariably put forward for a fairly narrowly delimited range of expressions, such as proper names or natural kind terms: such theories attempt to make informative generalizations about the common features which our dispositions to apply and interpret members of a given type of expression are sensitive to.

We thus reach the following explication of the reference relation:

[12] It should be noted that deRosset is arguing against the view that deference should depend on explicit metalinguistic or metacognitive intentions. Our version of meta-internalism merely requires that deference be grounded in deferential dispositions; it is an open question whether such deferential dispositions require a theory of mind unavailable to three-year-olds. Our point here is merely that *if* three-year-olds are not capable of such deferential dispositions, this is a good ground for denying that they are not fully competent in the deferential use of proper names. Whether the appropriate deferential dispositions require a theory of mind is a crucial question, but one which we will not attempt to resolve here.

[13] From here onwards, we will be assuming that languages are shared—that is, that in fact groups of speakers have dispositions to react to each others’ usage of words and co-ordinate their usage.

[14] Construing the reference relation in dispositional terms makes our account, of course, vulnerable to Kripkensteinian worries [Kripke 1982]. Unsurprisingly, we will not solve Kripke’s Wittgenstein’s rule-following paradox in this paper. Suffice it to say that we are not convinced that the paradox is a knock-down argument against dispositionalism; we believe that a version of dispositionalism roughly along the lines that Philip Pettit suggests in his “Ethocentric” solution [1996]) can withstand Kripke’s Wittgenstein’s attack.

[15] Which would involve, for example, that they are disposed, under favourable circumstances, to choose the expression in question when they intend to talk about Barack Obama, and are disposed to take utterances containing the expression in question to be about Barack Obama.

> *Reference*: A token expression e refers in language L to object o iff (i) e is standing in the R-relation to o and (ii) competent speakers of L are disposed to interpret objects (of the type of o) to be the referents of expressions (of the type of e), if they believe these are connected by the R-relation.

Theories of reference for particular types of expressions attempt to specify the relevant R-relation. For example, one theory of the reference of proper names claims that the R-relation is that of uniquely satisfying the description that the speaker (of those tokens) associates with the name. Another theory claims that the R-relation for proper names is that of (being an object) standing at the other end of the causal-historical chain of reference-borrowing that leads up to the (token of the) proper name. Which of these specifications (if any) of the R-relation is correct for English is determined by the dispositions that speakers of English possess to apply and interpret proper names.

A note might be in order about the final phrase 'if they believe these are connected by the R-relation'. Of course, the whole explication should not be understood as requiring that the speakers have any beliefs about what is constitutive for a term to refer to an object. Thus the beliefs that a speaker has "about the R-relation" do not need to represent this relation to her as a reference relation. It is just supposed to mean that a speaker might believe, for example, that H_2O is the unique underlying microstructure of water and therefore will intuitively use and interpret 'water' as referring rigidly to H_2O. In fact, for 'water' to refer to water, competent speakers do not need to have any beliefs about 'water' and water. It suffices that their disposition to use the term should be sensitive to information about water and 'water', such that if they were aware that 'water' is standing in the R-relation to water, they would interpret 'water' as referring to water.

This might also help to clarify how our account differs from other first-order response-based theories, such as Jackson [1998] and Chalmers and Jackson [2001], and Gertler [2002]. These accounts typically imply that under ideal epistemic circumstances, if confronted with a t-neutral description of the totality of facts, competent speakers are in a position to know what the reference of t is, for any term t in their repertoire.[16] Let us call this the "scrutability of reference thesis". Although the scrutability of reference thesis is compatible with our explication *Reference*, it is not implied by it.[17] To see this, one should note that our dispositions do not need to be transparent to us. It is conceivable that, when considering counterfactual states of affairs, we assume that we would interpret a term's reference in accordance with relation R_1, but were such states of affairs to become actual, we would in fact interpret its reference according to another relation, R_2. Thus, even if our usage is in fact sensitive to that information about the totality of facts, this does not imply that we have a safeguarded *a priori* access to facts about the outputs of that disposition.[18] Moreover, we can (arguably) imagine situations, consistent with *Reference*, where the reference of t is not *a priori* scrutable from a t-neutral description of the totality of facts, even though the R-relation that determines t's reference is scrutable. For example, for some terms in my repertoire I might just defer to the dispositions of experts to use a term, but there might not be a t-neutral

[16] For a characterization and critique of response-based theories in this sense, see deRosset [2011].

[17] And neither is *Reference* implied by the scrutability of reference thesis. The scrutability of reference thesis could be true, for example, if it would always take hard intellectual work to figure out what the reference of an expression is given a totality of facts, while the dispositions to use and interpret that expression would be totally unaffected by the information about the totality of facts.

[18] Of course, to some extent it is plausible to assume that we have such access, and to that extent considering thought experiments in order to determine what dispositions we have makes sense. We say more about this below in 6.3.

characterization of the expert's dispositions to apply those terms that I would recognize as the relevant dispositions for determining reference.

6.2. 'Reference' is Not a Natural Kind Term

On the moderate meta-internalist picture, reference is thus a dispositional or functional property. This is in stark opposition to (at least most versions of) meta-externalism and radical meta-internalism, which seem to be treating reference as a natural kind, the underlying nature of which theories of reference attempt to discover. We think, however, that there is no underlying nature to be discovered: there is good reason to deny that 'reference' is a natural kind term. However, this reason is partly dependent on moderate meta-internalism. Accordingly, to use this to argue *for* moderate meta-internalism would be to beg the question. But we take ourselves to have already given good grounds for accepting moderate meta-internalism—the following argument merely illustrates how and why 'reference' should not be considered a natural kind term, given moderate meta-internalism.

Suppose moderate meta-internalism is true. It follows, among other things, that the question of what kind of a property a given expression "aims to refer"[19] to is determined by factors internal to us. Suppose, further, that first-order externalism holds for natural kind terms. Such first-order externalism is, then, made true by the kinds of burden-shifting dispositions mentioned earlier. 'Water' aims to denote a natural kind because our dispositions to apply and interpret the term 'water' are sensitive to how the actual world turns out to be: we evaluate applications of 'water' partly on the basis of whether the substance that 'water' is applied to shares the underlying structure of the watery stuff of our local acquaintance.

All this is reflected in the fact that 'water' is twin-earthable: it is precisely because our dispositions shift part of the burden to the empirically discoverable external facts that we can imagine watery stuff that is not water. Indeed, twin-earthability appears to be (part of) what makes 'water' a natural kind term (that is, a term that aims to denote a natural kind, whether or not it succeeds). Other terms, such as 'bachelor', are not twin-earthable—they do not even aim to denote natural kinds. Even if it turned out, miraculously, that all and only bachelors actually have some empirically discoverable microstructure, that would not make bachelorhood into a natural kind: in other possible worlds 'bachelor' would apply on the basis of gender, age, and marital status, not microstructure.

What about 'reference', then? Is it twin-earthable? Herman Cappelen and Douglas G. Winblad argue that 'reference' indeed is twin-earthable, and that this has consequences for the methodology used in philosophy of language. They in fact try to establish that 'reference' is like a natural kind term in the sense that it is external facts, outside the head, that determine what it refers to, such that thought experimentation about 'reference' should be abandoned, being an unreliable methodology in the case of terms that have their extension fixed externally. As noted above, Cappelen and Winblad's argumentation is a prime example of meta-externalism.

Cappelen and Winblad ask us to imagine that the actual world is a Kripke-world, where a Kripkean causal theory of reference is true, while Twin-Earth is an Evans-World, a world where Gareth Evans' theory of reference is true. Now consider the case of Alice, the Earthian, and Twin-Alice, who are both ignorant about theories of reference and in particular ignorant about what theory of reference is true, but both are nevertheless familiar with the word 'reference'. Since Alice and Twin-Alice are possibly psychologically indistinguishable (or so Cappelen and Winblad argue), this should establish externalism about 'reference'.

[19] The notion of (fallibly) "aiming to denote a natural kind" is borrowed from McLaughlin and Tye [1998].

This argument, however, is not suitable to establish that ‘reference’ is not a functional kind term (and that its reference is externally determined). We can consider an uncontroversial functional kind term like ‘hygrometer’ and assume that Berta and Twin-Berta know that hygrometers are devices for measuring humidity in the air. However, they are both ignorant of the mechanisms that hygrometers in their world use to determine the level of humidity in the air. We may assume that Berta and Twin-Berta inhabit worlds that are different only in the way that hygrometers are built. Now Berta and Twin-Berta might be in phenomenologically identical situations with respect to what they call ‘hygrometer’, but the extensions of ‘hygrometer’ on Earth and Twin-Earth would be different. Should that mean that the reference of ‘hygrometer’ is determined externally?

Of course not. ‘Hygrometer’ refers to all and only those devices that measure humidity in the air. This is what Berta and Twin-Berta “have in their heads” and this is what determines the extension of ‘hygrometer’ in their worlds. But then something must be wrong in Cappelen’s and Winblad’s reasoning, when they argue that:

> The extension of ‘reference’ for Alice and Twin Alice differ; but what is in their minds is the same. If this situation shows externalism to be true about ‘water’ it should do the same for ‘reference’.
>
> [Cappelen and Winblad 1999: 339]

We believe that it is not difficult to identify the flaw here. Cappelen and Winblad mistakenly assume that this part of the story about Oscar and Twin-Oscar and the term ‘water’ already establishes externalism about ‘water’ by itself. To see that it does not do this, we may assume that ‘water’ is not a natural kind term, but a functional kind term. In that case, Oscar and Twin-Oscar, both ignorant about the microstructure of water, sitting in front of a glass of water, might both be in the same psychological state when saying ‘there is water in the glass on the table’, even though the extension of ‘water’ in their worlds is different. This does not show at all that ‘water’ has an extension that is externally fixed. What Cappelen and Winblad overlooked is that the thought experiment by Putnam does not stop there. Putnam asks us to imagine what would happen 150 years later, after the discovery of water’s microstructure on Earth, when a Twin-Earth mission visits Twin-Oscar’s planet and inquires what substance it is that Twin-Earthlings call ‘water’. As Putnam’s thought experiment suggests, we Earthlings would not in that case say that there is water on Twin-Earth. We would treat ‘water’ as a rigidly referring natural kind term. Thus: *what is in the head of Oscar and Twin-Oscar does not fix the extension for every given possible world*, hence the meaning of ‘water’, being a function from possible worlds to extensions, is not in the head of either Oscar nor Twin-Oscar.

Can we say the same about ‘reference’? It seems to us that we cannot. The argument is largely similar to the one above against meta-externalism. Suppose that meta-internalism is correct and that what determines how terms refer is the dispositions of speakers to use and interpret terms in certain ways. Now imagine two worlds; Carla’s world in which the dispositions of speakers are such that a descriptivist theory of names is true for their usage of names, and Twin-Carla’s world in which a causal-historical theory of names is what accords with those dispositions in her linguistic community. Carla and Twin-Carla, when talking about reference, might be using the term ‘reference’ with two different extensions. However, when Carla graduates in philosophy of language and boards a space ship to visit her soulmate on Twin-Earth, she would not report back to Earth (after reading up on Twin-Earth philosophy of language) that on Twin-Earth names mysteriously are not used to refer to their bearers, but that the way that reference of names works in Twin-Earth-English is just very different from the way that it works in Earth-English. If Carla and Twin-Carla only know the biconditional we called "*Reference*", then what is in their head, even if they are ignorant about what the R-relation is in their linguistic community, is already sufficient for determining the intension of 'reference'.

There is, however, a second sense in which the meaning of 'reference' might be fixed "externally" to the mind of a speaker. It is worth looking into this to uncover a second confusion in Cappelen's and Winblad's argumentation. Take again an uncontroversial functional kind term, like 'hygrometer' and assume this time that Paula and Twin-Paula are similarly ignorant about what hygrometers are *for*. In fact, in Paula's world they are used to measure humidity in the air, while in Twin-Paula's world 'hygrometer' refers to devices that measure the density of liquids. Paula's and Twin-Paula's fathers are, however, in the hygrometer-business on their respective planets. Paula and Twin-Paula both know that their garages are filled with boxes of hygrometers, that hygrometers are not selling as well anymore as they did before the Chinese started to mass produce them and flooded the market, etc. Thus, although Paula and Twin-Paula are ignorant about the function of "hygrometers", they might, to some extent, be competent with using the term and just defer to experts (like their fathers) when it comes to determining what exactly 'hygrometer' refers to. Thus, also with respect to functional kind terms, it might be plausible to think that their meaning is fixed in a social external way. Since 'hygrometer' is not a word of ordinary English or ordinary Twenglish, this is likely to be the case.

What about 'reference'? Also for the case of 'reference' we believe that it is a theoretical term that an ordinary speaker does not need to be able to command in order to be linguistically competent. Thus, merely on the basis of linguistic competence, one does not automatically know that *Reference* is true. It is easy to imagine Michael and Twin-Michael who fail to recognize that this is how the extension of 'reference' is determined, but whose usage of the term 'reference' still manages to hook onto the reference relation thanks to the division of linguistic labour. Does that not still establish Cappelen's and Winblad's conclusion, viz. that thought experimenting and intuition-probing about 'reference' should be unreliable since there is no reason to believe that our intuitions concerning 'reference' are sufficiently guided by the relevant individuation conditions of 'reference'?

Of course, one can agree that 'reference' is a technical term. But what is being tested, for example, in Putnam's thought experiment is not our intuitive conception of the term 'reference'.[20] When we ask whether, by our intuitions, the crew should report back to Earth that there is no water on Twin-Earth, we are not interested in our intuitions regarding technical vocabulary. Consider an analogous case with the notion of 'folk psychology'. 'Folk psychology' is, arguably, also a theoretical term. It features in a theory, let's call it 'FPT' for 'Folk-Psychology-Theory', that explains how human beings make sense of, predict, and explain each other's behaviour. In FPT, it is assumed that human beings possess a tacitly represented theory, viz. folk psychology, which is activated in their intuitive, dispositional ascription of beliefs, intentions and desires to other human beings.

Although 'folk psychology' is not a term that ordinary human beings are likely to have any interesting intuitions about, 'folk psychology' refers, if folk psychology exists, to a cluster of psychological dispositions to make certain intuitive ascriptions of beliefs, intentions, and desires in certain circumstances. Hence, even though it does not make much sense to study folk intuitions when it comes to the meaning of 'folk psychology' (because it is a theoretical term), it does nevertheless make a lot of sense to study folk intuitions when studying the nature of folk psychology, because folk psychology just is (supposedly) the psychological dispositional state of having certain kinds of intuitions.

[20] This is also overlooked by Noam Chomsky [2000] and Stainton [2006] when they argue in a similar way that thought experiments in philosophy of language are useless because 'reference' is a theoretical term.

We believe that the situation is essentially the same with respect to ‘reference’. ‘Reference’ is a theoretical term within philosophical semantics that plays a role in the overall semantic theory of explaining how the truth-conditions of sentences are determined by the semantic values of sub-sentential expressions. Again, it is not likely that ordinary speakers have any illuminating intuitions about ‘reference’ so understood. However, the nature of reference, if it exists, is again determined by a complex dispositional state of ordinary speakers to use and interpret expressions of certain types in certain ways. And it is intuitions as the outputs of these dispositional states that are “tested” in thought experiments about reference. In the next section we will try to say a bit more about what consequences this might have for the use of intuitions in the philosophy of language.

6.3. The Methodology of Theories of Reference

Given what we have said so far, it should be clear that intuitions elicited by thought experimentation can, on our view, be relevant to theories of reference. If meta-internalism is true, referential properties are wholly determined by our referential dispositions and practices, and it seems clear that thought experiments are a source of evidence for such dispositions. But it is less clear whether thought experiments are the *best* source of evidence for the kinds of dispositions that are relevant for the determination of reference. Moreover, even if thought experiments are to be used in arguing for and against theories of reference, it is not immediately clear *whose* dispositions we should primarily investigate—the philosophers’ or the laymen’s—and how. These are difficult questions, which we cannot hope to fully resolve here—and it may well be that the question of what is the *best* source of evidence cannot be settled *a priori*. But some tentative methodological conclusions of our view are worth pointing out.

If *Reference* is even roughly correct, the linguistic dispositions of all competent speakers will be relevant for the determination of reference. There is no reason to think that the dispositions of some subgroup of speakers (such as Western philosophers of language) should reflect some special expertise or insight into semantic relations. Accordingly, empirical evidence on the dispositions of non-philosophers could potentially be highly relevant to theories of reference, and cannot be dismissed out of hand. For example, should it turn out that Kripkean intuitions about the use of proper names are not widely shared outside the circle of Western philosophers of language, our confidence in the causal-historical theory of reference for proper names would be severely undermined. But we should be careful not to jump into dramatic conclusions prematurely: disagreement regarding the kinds of thought experiments standardly used in philosophy of language does not immediately indicate disagreement in the kinds of dispositions relevant for the determination of reference, i.e. speakers’ dispositions to interpret the utterances of others in certain ways, and to produce utterances with certain proper names given certain communication intentions. It may well turn out that such differences can be tracked down to other factors.

In particular, as Genoveva Martí [2009] stresses, the results reported by Machery et al. [2004], concerning non-philosophers’ intuitions in response to Kripkean thought-experiments about proper names, are highly problematic as evidence for how speakers actually use and interpret proper names. Machery et al.’s subjects were asked to provide fairly theoretical judgments about whom characters in fictional scenarios were using proper names to refer to. Giving answers to such questions requires a great deal of generalization over a range of possible particular applications and interpretations of proper names, which gives rise to additional sources of error. We should bear in mind that thought experimentation is a philosopher’s tool: it is not inconceivable that philosophers should, because of their training and experience, be more skilled at adequately reporting and forming generalizations on the basis of their linguistic dispositions. Of course, the extent to which this *is* the case is an empirical question, but this possible explanation of the apparent divergence in intuitions cannot simply be dismissed.[21]

Nevertheless, it does follow from *Reference* that empirical work on ordinary speakers' use- and interpretation-intuitions can potentially be highly relevant for theories of reference. But asking ordinary speakers to engage in thought-experimentation, especially in the prevalent form which makes use of fictional (and sometimes far-fetched) scenarios and forced-choice questions, is not likely to be the best possible source of evidence for deciding which theory of reference correctly captures the everyday usage of linguistic expressions—we should rather try to test for speakers' linguistic dispositions without having to mention the notion of reference, and without inviting the test subjects to meta-reflect on their referential and interpretational practices. More informative data would have to concern the use and interpretation of referring terms more directly, and such studies would have to more clearly distance themselves from the philosophical thought-experiment paradigm than current studies have done.

7. Conclusions

We have argued that another internalism–externalism distinction should be drawn, in addition to the familiar one concerning what determines the extension of linguistic expressions, and that this distinction between meta-internalism and meta-externalism is conceptually independent of the first-order distinction. Making the meta-level distinction is, we think, crucial in clarifying questions concerning what the proper goals and methods of a theory of reference are (and should be).

We have also argued for a particular position on the meta-level, moderate meta-internalism. This view seems to make the best sense of the aims of theories of reference, while also providing a partial justification for the intuitive methodology that theorists of reference have traditionally relied on. At the same time, however, moderate meta-internalism entails that empirical evidence about speakers can be highly relevant for theories of reference. In the last section we argued that existing experimental setups are unlikely to provide very good grounds for substantial conclusions in the theory of reference. If moderate meta-internalism is accepted, however, a clearer picture emerges of what kinds of states in speakers should be studied: moderate meta-internalism can thereby be used to guide more fruitful empirical work in philosophical semantics in the future.[22]

Department of Philosophy, University of Tartu, Estonia
Department of Philosophy, Norwegian University of Science and Technology, Trondheim, Norway

REFERENCES

Burge, T. 1979. Individualism and the Mental, *Midwest Studies in Philosophy* 4/1: 73–122.

Bezuidenhout, A. L. 2006. Language as Internal, in *The Oxford Handbook of Philosophy of Language*, ed. E. Lepore and B. C. Smith, Oxford: Oxford University Press: 127–139.

[21] See Machery et al. [2009] for a reply to Martí. See also Sytsma and Livengood [2011] for a recent critical response to the initial study by Machery et al.

[22] The material in this paper was presented at various conferences and workshops during the last months. We would like to thank those audiences, two anonymous referees for this journal, Magdalena Balcerak Jackson, Brendan Balcerak Jackson, Michael Bishop, Jekaterina Denissova, Michael Devitt, Heidi Haanila, Nat Hansen, Sören Häggqvist, Edouard Machery, Genoveva Martí, John Perry, Anne Salminen, Barry C. Smith, Stephen Stich, and Åsa Wikforss for helpful comments. Research for this paper was supported by Estonian Science Foundation grants SFLFI11085E, SF0180110s08, and ETF9083, and the Norwegian Research Council grant 212841.

Cappelen, H. and Winblad, D. 1999. Reference Externalized and the Role of Intuitions in Semantic Theory, *American Philosophical Quarterly* 36/4: 337–50.

Chalmers, D. J. and Frank Jackson 2001. Conceptual Analysis and Reductive Explanation, *The Philosophical Review* 110/3: 315–60.

Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*, Cambridge: Cambridge University Press.

deRosset, L. 2011. Reference and response, *Australasian Journal of Philosophy* 89/1: 19–36.

Devitt, M. 1994. The Methodology of Naturalistic Semantics, *The Journal of Philosophy* 91/10: 545–72.

Devitt, M. 2006. *Ignorance of Language*, Oxford: Clarendon Press.

Devitt, M. 2009. On determining what there isn't, in *Stich and his critics*, ed. D. Murphy, and M. Bishop, Chichester: Blackwell: 46–61.

Devitt, M. 2011a. Experimental semantics, *Philosophy and Phenomenological Research*, 82/2: 418–35.

Devitt, M. 2011b. Deference and the use theory. *ProtoSociology* 27: 196–211.

Devitt, M. 2012. The Role of Intuitions, in *Routledge Companion to Philosophy of Language*, ed. G. Russell and D. Graff Fara, London and New York: Routledge: 554–65.

Devitt, M. and Sterelny, K. 1999. *Language and Reality: An Introduction to the Philosophy of Language*. Oxford: Blackwell.

Gertler, B. 2002. Explanatory Reduction, Conceptual Analysis, and Conceivability Arguments about the Mind, *Noûs* 36/1: 22–49.

Haukioja, J. 2009. Intuitions, Externalism, and Conceptual Analysis, *Studia Philosophica Estonica* 2.2: 81–93.

Häggqvist, S. and Wikforss, Å. 2007. Externalism and a posteriori semantics, *Erkenntnis* 67/3: 373–86.

Jackson, F. 1998. Reference and Descriptions Revisited, in *Philosophical Perspectives 12: Language, Mind, and Ontology*, ed. James E. Tomberlin, Oxford: Blackwell: 201–18.

Korman, D. 2006. What Externalists Should Say about Dry Earth, *Journal of Philosophy* 103/10: 503–20.

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

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.

Laurence, S. 2003. Is linguistics a branch of psychology?, in *Epistemology of Language*, ed. A. Barber, Oxford: Oxford University Press: 69–106.

Ludlow, P. 2003. Externalism, Logical Form, and Linguistic Intentions, in *Epistemology of Language*, ed. A. Barber, Oxford: Oxford University Press: 399–414.

Ludlow, P. 2011. *The philosophy of generative linguistics*, Oxford: Oxford University Press.

Machery, E., Mallon, R., Nichols, S., and Stich, S. 2004. Semantics, cross-cultural style, *Cognition* 92/3: B1–B12.

Machery, E., Oliviola, Christopher Y., and De Blanc, Molly 2009. Linguistic and metalinguistic intuitions in the philosophy of language, *Analysis* 69/4: 1–6.

Mallon, R., Machery, E., Nichols, S., Stich, S. 2009. Against arguments from reference, *Philosophy and Phenomenological Research* 79/2: 332–56.

Martí G. 2009. Against semantic multi-culturalism, *Analysis* 69/1: 42–8.

McLaughlin, B. and Tye, M. 1998. Is Content-Externalism Compatible With Privileged Access?, *Philosophical Review* 107/3: 349–80.

Pagin, P. and Westerståhl, Dag 2010. Compositionality II: Arguments and problems, *Philosophy Compass* 5/3: 265–82.

Pettit, P. 1996. *The Common Mind*, New York: Oxford University Press.

Putnam, H. 1975. The Meaning of 'Meaning', *Minnesota Studies in the Philosophy of Science* 7: 131–193.

Searle, J. 1958. Proper Names, *Mind* 67/122: 166–73.

Segal, G. 2001. Two theories of names, *Mind & Language* 16/5: 547–63.

Stainton, R.J. 2006. Meaning and Reference: Chomskyan Themes, in *The Oxford Handbook of Philosophy of Language*, ed. E. Lepore and B. C. Smith, Oxford: Oxford University Press: 913–40.

Stainton, R. J. 2010. The role of psychology in the Philosophy of Language, in *Routledge Companion to Philosophy of Language*, ed. G. Russell and D. Graff Fara, London and New York: Routledge 525–532.

Stich, S. 1996. *Deconstructing the Mind*, New York and Oxford: Oxford University Press.

Sytsma, J. and Livengood, J. 2011. A new perspective concerning experiments on semantic intuitions, *Australasian Journal of Philosophy*, 89/2: 315–32.