Variation in Natural Kind Concepts

Daniel Cohnitz and Jussi Haukioja

Abstract

Since Kripke's and Putnam's work in the 1970's, most philosophers have assumed that our natural kind concepts are externally individuated. However, both psychologists and philosophers have questioned this assumption, partly on empirical grounds. One strand of such criticisms is that there may well be systematic variation in how subjects apply natural kind terms either across persons or across times, and that we should therefore be prepared to accept that natural kind concepts are not as universally shared, or temporally stable, as many philosophers have been assuming.

Yet, it is far from clear exactly what kind of variation in subjects' application of natural kind terms would seriously cast doubt on the view that natural kind concepts *are* externally individuated. In this paper, we will take a detailed look at this question, building on the dispositionalist meta-internalist perspective in meta-metasemantics that we have developed in earlier work. We will look at what kinds of speaker responses are relevant for evaluating externalist views, and review existing empirical research on natural kind concepts against this background. We will argue that the existing studies do not call for dramatic revisions to the externalist mainstream, and conclude by exploring some possible new directions for the experimental study of natural kind terms and concepts.

1. Introduction

On a textbook internalist conception of meaning, concepts[1] seem to be very unstable. If a term's meaning is determined by the speaker's beliefs, associated (bundle of) descriptions, or dispositions to apply the term to objects in the world, it would seem that conceptual variation is quite commonplace, between speakers as well as between different temporal slices of the same speaker. We change our beliefs about all kinds of things relatively frequently, and the same holds for any descriptions we might associate with terms. Already Frege noticed that different speakers well might, and often probably do, associate different descriptions with a name. We know from empirical research that people vary in their dispositions to apply terms, between cultures, within cultures, and even interpersonally (for the latter, cf. Hampton & Passanisi 2016).

1 We follow the convention of philosophers of language and will use 'concept' for the meaning of predicate expressions, which determine the reference (extension) of these expressions. If two expressions differ in extension, they must also differ in meaning. Because of that latter connection, philosophers often discuss variations in extension when discussing conceptual variation in this sense. By identifying concepts with meanings, it is left open whether they are psychological entities or closely related to such. Note that in psychological literature "concepts" are often by definition psychological entities.

This kind of instability leads to well-known puzzles. How can speakers ever mean the same by their terms, if meaning is subject to such variation? Doesn't that show that we speak past one another almost all the time? How is communication then possible? And, relatedly, how is rational theory change and scientific progress possible, if the terms of scientists in the past had a different meaning from those of scientists today?

Externalism about meaning promises a solution to these problems. Meaning is not determined by our variable psychological states, instead it is determined by external factors; perhaps facts about the underlying kinds and their essences or properties[2], perhaps facts about the linguistic community as a whole. Facts of the latter kind might prevent us from talking past one another, facts of the former kind might moreover guarantee that we use terms with the same meaning even if our beliefs about their referents change radically over time. Externalism of this kind, we will call *first-order externalism* (and it contrasts with first-order internalism as sketched above). But is first-order externalism true? How can we find out? What kind of facts would make it true? What kind of facts, in general, make metasemantic theories true or false?

As we will argue in the next section, one can have two radically different views about this. One may think that the metasemantics of natural kind terms is determined by certain dispositional states of speakers (a view that we will call "meta-internalism") or hold that it is determined by speaker-external facts, such as certain metaphysical facts or perhaps supra-individual facts about the linguistic community as a whole (a view that we will call "meta-externalism"). We will argue in the next section that the latter view is implausible. Metasemantics must be tied to speaker dispositions of a certain kind in order to allow our metasemantic theory to be of explanatory value.

As we will argue in section 3, this view has methodological implications. If certain dispositional states of speakers determine metasemantics, then we should inquire what these dispositions are, in order to find out whether the externalist metasemantic view that we sketched above is right and can indeed solve the puzzles described. Investigating these dispositions can be done empirically.

2 It is sometimes assumed that externalism is already committed to the existence of essences, such that if there are no properties that could play the role of essences, then externalism is false (c.f. Häggqvist & Wikforss forthcoming). But this is a misunderstanding. First of all, essences might be conferred, and thus not require a metaphysical foundation (Sveinsdottir 2008). Secondly, the dispositions of speakers to defer to whatever gives rise to the stereotypical properties of a natural kind in the actual world, are probably highly flexible and might well settle for bundles of properties, or disjunctions (or for whatever contemporary scientists, if well-informed, would consider the most salient realizers for the assumed stereotypical properties, or for some of them). The point of externalism is precisely that the properties (be they essential, intrinsic, or highly disjunctive) that do determine the extension of a natural kind term are not necessarily known to a speaker who uses a natural kind term with that meaning.

However, the kinds of experimental studies that have so far been performed to evaluate the truth of externalism are insufficient for deciding the matter. In section 4 we will discuss how existing methods could be improved upon.

2. Two stories about meta-metasemantics

The standard Twin-Earth argument for externalism is very well-known. We introduced the term 'water' in order to talk about a certain liquid we find in rivers, lakes and oceans. The term then applies in our world (but also in other possible worlds) to liquids that share the same chemical composition as that of water. Thus, on Twin-Earth, where the "watery stuff" is XYZ, we find no water.

Compared to the externalist story about proper names, we have now *two* places where external factors determine the reference of a term like 'water'. First, just as in the case of names, there is the causal-historical chain of usage that connects our usage of the word with ourworldly water. Just as with names, we don't need to know anything about this chain in order for 'water' to refer to water.

But the extension of 'water' is also determined by the chemical composition of water.[3] Whatever shares this microstructure is water, even if we don't know what that microstructure is and even if we don't know that there is such a thing as chemical composition. As Putnam argued in the case of gold:

> It is possible (and let us suppose it to be the case) that just as there were pieces of metal which could not have been determined not to be gold prior to Archimedes, so there were or are pieces of metal which could not have been determined not to be gold in Archimedes' day, but which we can distinguish from gold quite easily with modern techniques. Let *X* be such a piece of metal. Clearly *X* does not lie in the extension of 'gold' in standard English; my view is that it did not lie in the extension of 'χρυϛος' in Attic Greek, either, although an ancient Greek would have mistaken *X* for gold (or, rather, χρυϛος). (Putnam 1975, 235)

But what makes it the case that 'gold' or 'χρυϛος' have their extension determined in this way? After all, not all terms work this way. The extension of 'bachelor' is not determined by

3 Or to whatever it is that accounts for the stereotypical properties of water in the actual world. See the previous footnote for a brief explanation why externalism is not committed to the view that the extension of 'water' is determined by some unique microstructure. The examples we use pretend that there is such a unique microstructure, for ease of exposition.

microstructure and neither is the extension of ‘pencil’, or countless other general terms. What is it about natural kind terms that makes them behave differently?

The possible options for answering that question can also be best illustrated with a case from Putnam’s “Meaning of ‘meaning’”:

> Imagine that we someday discover that pencils are organisms. We cut them open and examine them under the electron microscope, and we see the almost invisible tracery of nerves and other organs. We spy upon them, and we see them spawn, and we see the offspring grow into full-grown pencils. We discover that these organisms are not imitating other (artifactual) pencils - there are not and never were any pencils except these organisms. It is strange, to be sure, that there is lettering on many of these organisms - e.g. BONDED *Grants* DELUXE made in U.S.A. No. 2 - perhaps they are intelligent organisms, and this is their form of camouflage. (Putnam 1975, 242)

As we said above, the extension of ‘pencil’ seems to be determined by a certain functional description. Pencils are artefacts that one can write with in virtue of a “narrow, solid pigment core inside a protective casing which prevents the core from being broken or leaving marks on the user’s hand during use” (Wikipedia). Could we then discover that pencils are organisms?

Putnam’s own answer is that this is indeed an epistemic (although not metaphysical) possibility, because “we intend to refer to whatever has the same nature as the normal examples of the local pencils in the actual world” (Putnam, 243). But what if this is a misdescription of our intention? What if we indeed intend to refer to whatever serves the same function thanks to a similar mechanism (by having a narrow, solid pigment core inside a protective casing which prevents the core from being broken or leaving marks on the user’s hand during use, for example)? Could the fact that what we call ‘pencils’ all turn out to be organisms *reveal* that ‘pencil’ does and always did refer to whatever shares the same DNA with pencils (whether or not you can write with them in other possible worlds), or would it still be the case that ‘pencil’ refers in this and other possible worlds to objects with the help of which one can write on paper and similar surfaces? In other words, is the metasemantic question of whether an expression is a natural kind term or a functional kind term a matter of our intentions and dispositions to use the term, or a matter of metaphysics?

One can be an *externalist* or an *internalist* about what determines metasemantics. In earlier work (Cohnitz & Haukioja 2013) we labelled these opposing positions "meta-internalism" and "meta-externalism". The two views are defined as follows:

> *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 the individual psychological states of S at the time of U.
>
> *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.

'How a linguistic expression refers...' is our abbreviation for the different metasemantic options: is the expression a natural kind term (such that its extension in other possible world is determined by having the same nature as the things in its extension in the actual world) or is it an artefact/functional kind term (such that its extension in other possible worlds is determined by having the same function in these worlds as the things in its extension in the actual world), and so on.

We take it that most authors, also those who favour a first-order externalist view on concepts, subscribe to meta-internalism. The typical story is that it is *because* we are disposed to defer to underlying microstructure or the knowledge of experts in our usage of a term, that our term refers via this microstructure, or according to that expert judgment. We have just seen this in the quote by Putnam; he seems to hold that it is due to our "intentions to refer to whatever has the same nature as the normal examples of the local pencils in the actual world" that 'pencil' refers to a natural kind, if what we thought were pencils turned out to be organisms. Thus, the question of whether 'pencil' is a natural kind term or a functional kind term (or whether it has a more complex, conditional structure) is then a question of whether speakers indeed have those dispositions or intentions. Such questions should then, in principle, be accessible to empirical investigation.

2.1 Two examples for meta-externalism for kind terms

Since meta-internalism seems to be relatively common, one might wonder whether there are actually any instances of meta-externalism. We will discuss two examples. On both examples, the deferential dispositions of speakers come out being largely irrelevant for the metasemantics of

natural kind terms. Consequently, it becomes also less clear how one should go about determining the actual semantics for such terms.

2.1.1 Reference Magnetism

Hilary Putnam has famously argued that we can find for any global theory infinitely many interpretations that satisfy that theory. This is Putnam's model theoretic argument against "global descriptivism". The theoretical role of an expression, even in a global theory, doesn't fix the extension for that expression. Moreover, Putnam added, any constraint other than truth (i.e. satisfaction of the theory by a model) would be just another piece of theory, hence not a way out of the embarrassment of riches. David Lewis noticed in his "Putnam's Paradox" (Lewis 1984) that the last inference doesn't hold. Putnam's argument seems to overlook the fact that there is a difference between a theory satisfying a constraint and the constraint holding because we accept a theory stating it:

> [An additional constraint] C is not to be imposed just by accepting C-theory. That is a misunderstanding of what C is. The constraint is not that an intended interpretation must somehow make our account of C come out true. The constraint is that an intended interpretation must conform to C itself. (Lewis 1984, 225)

But Lewis was also aware that this observation alone doesn't help. The distinction between satisfying C-theory vs. conforming to C can only be made if constraint C is not established merely by stipulation, by our referential intentions:

> The main lesson of Putnam's Paradox, I take it, is that this purely voluntaristic view of reference [viz. that whatever theory of reference is true, it is true because of our referential intentions] leads to disaster. If it were right, any proposed constraint would be just more theory. Because the stipulation would be something we say or think, something we thereby add to total theory.
>
> Referring isn't just something we do. What we say and think not only doesn't settle what we refer to; it doesn't even settle the prior question of how it is to be settled what we refer to. Meanings---as the saying goes---just ain't in the head. (Lewis 1984, 226)

As the last sentence of the quote already indicates, Lewis concluded from the observation that the constraint C (whatever it is) shouldn't hold just because we intend it to hold. The constraint must be something that is entirely independent of us and what's in our heads.

Lewis then suggests that the constraint is externally provided by the objective naturalness of certain elite properties. These elite properties are, for Lewis, fundamental physical properties, like mass, charge, quark colour and flavour. Obviously, most of our words do not refer to these, but Lewis believes that they do refer to derivatively eligible referents that are connected to the elite properties via chains of definitions. Properties that stand in such chains to the elite properties of fundamental physics are *reference magnets*: because they carve nature at the joints better than rival interpretations, they become the referents of our terms. Moreover, they become the referents of our terms, independently of our intentions concerning what our terms should refer to:

> It is not to be said that our theorising makes the joints at which the world is to be carved. That way lies the 'just more theory trap'. Putnam would say: `"very well, formulate your theory of 'objective joints in nature' [...] and stipulate if you will that your referents are to be 'eligible'. But total theory with this addition goes the way of all theory, it is satisfiable with the greatest ease in countless ways.'" [...] No: the proposed constraint is that referents are eligible, [...] not that the referents of 'cat' etc. are to be included among the referents of 'eligible'. (Lewis 1984, 228)

Now, of course, our intention *might* be that 'cat' refers to an eligible referent, like a natural kind, but the fact that 'cat' refers to a natural kind is not because of these intentions, it is solely because there is a natural kind and the assignment of this kind to the word 'cat' is part of an interpretation that maximises the eligibility of referents overall. It is clear why this is an instance of what we have defined as meta-externalism. As Lewis says in the quote we gave above: "what we say and think not only doesn't settle what we refer to; *it doesn't even settle the prior question of how it is to be settled what we refer to*" (Lewis 1984, 226, emphasis DCJH).

2.1.2 Reference Communitarianism

A second view that would count as meta-externalist takes less issue with the meta-internalist idea that semantic properties (like the metasemantics of an expressions) must ultimately be grounded in the way speakers use a language (which seems to be a very common idea anyway), but rather with the *individualism* that we built into our definition. We said that it should be a matter of the

dispositions/intentions of a particular speaker at the time of a particular utterance that determines the metasemantics of the expressions in that utterance. This is denied by views which would accept that these intentions occur at other times (for example, at some point in the past when the speaker first acquired the relevant terms, cf. Devitt 2011). It is also denied by views which would hold that semantic facts supervene on the usage of expressions in the linguistic community as a whole *and* hold that the entrance ticket to the linguistic community is obtained through, for example, linguistic interaction with that linguistic community (Williamson 2007).

Thus, if the usage in the linguistic community is such that 'pencil' would turn out to be a natural kind term, if we made the discovery that Putnam describes, then the unwillingness to change usage and retract the former claims about pencils being artifacts, etc. of any individual speaker who previously engaged linguistically with the linguistic community that Putnam envisages in his example, wouldn't betray that speaker's deviant meaning of 'pencil' in her idiolect but simply display that speaker's ignorance of the right metasemantics of that term.

2.2 Against meta-externalism

Language, especially when used in linguistic communication, helps us to coordinate our activities. It allows us to share knowledge about the world, on the basis of which we can then take action, but it also allows us to make and share plans that we can then carry out together. That our words express the concepts they do will somehow have to be a systematic part of any explanation of the role of language in this. The problem with meta-externalism is that it makes that systematic contribution superfluous.[4]

The impact of language on our joint plans and activities (but also on our individual beliefs) is (at least) a matter of our dispositions to react to information of certain kinds. Let's assume that the relevant dispositions of a linguistic community *L* are in harmony. Their linguistic interaction, their dispositions to correct and change their usage of expressions in the light of new information is *as if* the expression *e* in their language was, say, a functional kind term, and had functional kind *f* as its referent. Let us call *f* the "schmeferent" of *e*. Let us assume though that some meta-externalist story is true for *e*. And, in fact, the dispositions of our harmonious linguistic community are actually out of step with *e*'s actual referent. Let us assume, for example, that, in fact, *e* is a natural kind term and the natural kind *n* is the referent of *e*.

4 The argument below is presented in much more detail in Cohnitz & Haukioja 2013.

If we want to explain how the speakers of our hypothetical linguistic community manage to coordinate their plans and actions with the help of *e*, we will need to talk about the schmeference of *e*. There will be no reason to bring in the reference of *e*. The latter will only be reasonable if the relevant dispositions of the speakers of the linguistic community in question are in pre-established harmony with the meta-externalistically determined facts about reference. But even when they are, these latter facts will not contribute to the explanation of the coordination achieved through linguistic communication.[5]

This is *always* so on the meta-externalist picture. Because the relevant meta-externalistic facts are in the past, in the future, or outside your head, they might not have any impact on how you coordinate with an expression (i.e. how you adapt your linguistic and non-linguistc behavior and how you update your beliefs in reaction to information phrased in terms of that expression). There might be more than one speaker for whom that's so, there might be more than two, and, ultimately, this might be so for all members of a linguistic community. In the latter, extreme case it is obvious that reference is not contributing to an explanation of linguistic communication. The cases in which meta-externalist facts *seem* to make a contribution are those in which schmeference and reference coincide.

2.3 Meta-internalism of the dispositionalist variety

A *dispositionalist* version of meta-internalism claims that the metasemantics of a referring expression, as used by a speaker, is supervenient on that speaker's dispositions to apply and interpret the expression in question.[6] The range of dispositions that will have to be included in the supervenience base is, however, quite wide. It will of course include the speakers' dispositions to apply the term to entities in the world, including her dispositions to apply and interpret the term in conditions considered as non-actual. However, the supervenience base will also have to include the speaker's dispositions to *revise* and *reconsider* her own applications of the term. There are at least

5 In his 2011, when defending his own version of reference magnetism, Ted Sider claims without argument that "clearly" the meta-externalistically determined referents of expressions would have to play a role in semantic explanations of thought, behavior and understanding (Sider 2011, 28). Not only isn't this "clearly" so, this just isn't so.

6 Above we talked a bit vaguely about dispositions and intentions and somewhat pretended that there is a way to characterize individual mental states in a purely internalist way. One does not need to endorse individualism or internalism about the mental in order to make sense of meta-internalism or the meta-internalism/externalism distinction. From our point of view it seems to make most sense to be a (first-order) externalist about the mental and drop the reference to intentions. The intention to use 'water' for water and 'pencil' for pencils will then be intentions to use either expression as a natural kind term, depending on how the world turns out to be. Meta-internalism and meta-externalism defined in terms of such intentions just collapse into one another.

Both notions should thus be defined in terms of dispositions to apply terms to objects and to correct previous applications in the light of new information. The characterisation of these dispositions can refer to the things in the world that they are sensitive to, it can also be applied to mental types (expressions in the lingua mentis, if you like).

three kinds of such corrective dispositions that will have to be considered. The first two kinds of corrective dispositions, when present, can make it the case that a given expression is to be given an externalist metasemantics. Dispositions of these kinds are the speaker's dispositions to revise her application and interpretation in response to empirical information about her surroundings, and in response to information about how relevant experts in one's speech community use the same expression. In our view, when a first-order externalist theory of reference is true of a term, it is *made* true by the fact that the relevant speakers are, in their application of the term, suitably sensitive to empirical information concerning such facts. The third kind of corrective disposition is more general, and has to be brought in to account for the distinction between correct and incorrect applications of terms, regardless of whether a first-order internalist or externalist view is true of the term in question.

The first kind of corrective disposition is sensitive to empirical information about the entities that putatively belong in the extension of a term. When speakers have such corrective dispositions, and they are systematic enough, physical externalism (or natural kind externalism) is true of a term. To illustrate, let us look at some examples. For some expressions, such as 'bachelor', our patterns and dispositions of application and interpretation appear to be unaffected by contingent features of the natural world around us. The properties that speakers – individually or collectively – associate with the expression are taken to be sufficient for determining whether the expression applies to a given individual or not. But with other expressions 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 the underlying structure of the substance that is causally connected in the appropriate way to our actual usage of 'water'. We also have dispositions to *re*-evaluate our application of such terms in the face of new empirical information about what the world is like.

To illustrate, suppose that we took a representative sample of bachelors, studied them empirically, and found out that every single one of them has a certain neural structure – call it N – while no individuals outside the sample have N. Would we then start to categorise people as bachelors, in the actual world as well as in other possible worlds, according to whether or not they have neural structure N or not? No we wouldn't: we would go on categorising people as bachelors, in the actual world as well as in counterfactual ones, according to their age, gender, and marital status. Or, consider another version: suppose that we found that, say, almost all unmarried adult males were found to have N, while a tiny proportion of them do not; at the same time, we find N present in a

very small number of married females. Would we then revise our categorisation of this small minority of men as bachelors and instead include the married females? No we wouldn't. But were we to make a similar discovery about the golden stuff, say that we find some of the metal we categorised as gold to have a different atomic number than 99.9% of the rest, while a small sample of a greenish looking metal turns out to have the same atomic number as the other 99.9%, we would, we think, revise our categorisation of the 0.1% and consider the greenish stuff as gold.

In the above, we have assumed that ordinary speakers in fact do have the kinds of dispositions that externalist thought experiments typically turn on. Of course, this is an empirical assumption, and one that can be (and has been) questioned. We will return to this issue in the next section. What matters here is that, on our dispositionalist meta-internalist view, such dispositions are what *makes* first-order externalism true (in contrast to meta-externalism, where they would have to be thought of as tracking, more or less successfully, an independently determined semantic reality). On our view, the truth or falsity of natural kind externalism turns on the existence and systematicity of precisely such dispositions.

The second kind of corrective disposition is sensitive to how other speakers use the expression in question - in particular, these are dispositions to defer to other, more expert, speakers. When present, and when systematic enough, such corrective dispositions make it the case that social externalism is true of an expression. For example, I can refer to elm trees with my term 'elm', even though I cannot tell them apart from beech trees, on the basis of my dispositions to defer to people who can actually tell elms apart from other trees (e.g. botanists or gardeners). Should I find that my classification of trees into elms and beeches differs from that of an expert, I would be disposed to immediately revise my earlier application of the terms and align my usage with that of the experts.

The third kind of corrective disposition is, as noted above, more general in that such dispositions need to be assumed to be present regardless of issues having to do with internalism and externalism. These are dispositions to reconsider and possibly take back one's own earlier applications of a term, based on closer consideration of the issue at hand, and not prompted by new information concerning one's surroundings or the linguistic usage of experts. All speakers make mistakes every now and then: not all instances of applying and interpreting expressions are correct. But theories of reference try to account for *correct* application and interpretation. One cannot, for example, argue against the view that the reference of 'bachelor' is determined by the associated properties of being unmarried, adult, and male, simply by pointing out that speakers are, in some situations, disposed to call

women, or married men, bachelors. In order for such an argument to have any force, it would have to be established that such applications are, in fact, *correct*.

As we have learned from the extensive debates concerning the Kripkensteinian problem of rule-following, it is by no means obvious how the distinction between correct and incorrect instances of application should be drawn. We cannot hope to settle this huge issue here, but it seems clear to us that corrective dispositions will have to have a central role here, either as constituents of correctness, or as paradigmatic evidence of incorrect use. Suppose, for example, that two speakers, A and B, both apply 'bachelor' to married men, on a few isolated occasions. One of the speakers, A, would instantly take back her classifications of these men as bachelors, were she to reconsider the cases (for example, in response to other speakers' challenging her use). Speaker B, on the other hand, would be unmoved, not at all disposed to take back her classifications, and insist that these men are bachelors, even though they are married. Obviously, in such a case, A's applications of 'bachelor' to married men would not pose a serious threat to the claim that the reference of 'bachelor', as used by A, is determined by the associated properties of being unmarried, adult, and male. Not so in the case of speaker B: if she really were to insist on calling some unmarried men bachelors, and not show any inclination to correct herself, we should conclude that the reference of 'bachelor', as B uses it, is not determined in the same way; she is not using the term with its standard meaning.

Obviously, a lot more should be said about all three kinds of corrective dispositions. The above sketch is, however, sufficient for our aims in this paper. The resulting view is one on which the reference of a given linguistic expression, as used by speaker S, is determined by how S is disposed to use the expression, how she is disposed to re-evaluate and revise her application in the light of new information about the world and about other speakers, and in the light of closer inspection of her own usage. Externalism and internalism concerning a given expression are then to be thought of views about which kinds of new information can lead to the relevant kinds of re-evaluations and revisions.[7]

7 This has a lot in common with the view recently proposed by Michael Johnson and Jennifer Nado, independently. On their view, "a linguistic expression *E* means some object, property, kind, relation, etc, *X*, in the mouth of speaker *S*, in virtue of the fact that *S* would be disposed to apply *E* to *X* if *S* had all the relevant information" (Johnson & Nado 2014, 81), where "relevant information" is said to consist of "the facts *F* that would, were *S* apprised of *F*, influence *S*'s dispositions to apply *E*" (ibid.). The difference between their view and ours appears to be mainly one of emphasis - we focus on *S*'s dispositions to change her dispositions to apply *E*, whereas they focus on the end-result of such (idealized) dispositions. A closer examination of the relationships between our views is, however, a topic for another occasion.

3. What empirical data has shown so far

Let us now return to the issue of conceptual variation. If meta-externalism were true, systematic variation in how speakers apply natural kind terms would, perhaps, not be a huge problem. For example, if one accepted reference magnetism, our natural kind terms would refer to the most natural of the plausible candidate referents (however exactly the class of such candidates is determined), and divergent applications could simply be dismissed. But we reject meta-externalism, so we need to meet the challenge head on. On our meta-internalist view, referential relations are determined by competent speakers' patterns of dispositions to apply the term to objects and to revise his or her applications in the light of new information. Thus, to appeal to first-order externalism about natural kind terms to explain successful communication and scientific progress, it has to be established that our dispositions in fact do support the purported referential stability of our natural kind terms. And, since the nature of our actual dispositions is clearly an a posteriori matter, such stability could be denied on empirical grounds, should it turn out that there is considerable variation in how we apply natural kind terms.

And indeed, this is precisely what some experimental studies claim to have shown. A number of theorists have claimed that, due to such variation, we should reject the (first-order) externalist view of natural kind terms that is predominant in philosophy of language, and replace it with a "representational change" theory according to which the reference of natural kind terms is context-dependent (Braisby et al 1996), a "hybrid" theory that includes both causal-historical and descriptive factors (Genone & Lombrozo 2012), or an "ambiguity" theory according to which natural kind terms can take on a causal-historical reading or a descriptive reading, depending on conversational setting (Nichols et al 2015).[8]

The empirical results that, according to these authors, establish that there is widespread variation in speakers' use of natural kind terms all have to do with test subjects' verbal responses to questions concerning imagined scenarios that are fairly closely modeled after the standard externalist thought experiments by Putnam and Burge. The questions take a variety of forms: in some cases, subjects were asked whether they agree with existence claims (Braisby et al 1996, Nichols et al 2015) or classifications of individuals as belonging to a natural kind (Braisby et al 1996) in response to Twin Earth -style scenarios, in others they were asked whether two speakers in Burge-inspired scenarios

8 A fourth study, by Jylkkä et al (2009) finds roughly similar results, but does not take them as evidence for an ambiguity view. We will briefly return to the results of this study in the next section.

were using terms (such as 'tyleritis', a term denoting a disease) co-referentially or not (Genone & Lombrozo 2012).

All studies show a similar pattern: in some situations, subjects were using and interpreting natural kind terms as a causal-historical theory would predict, while in others their usage was in line with a descriptivist theory. Moreover, none of the studies found subsets of subjects using terms exclusively according to one of the theories; all subjects appeared to be switching between the two patterns, even to the point that they accepted apparently contradictory pairs of sentences as true (Braisby et al 1996, Nichols et al 2015).

As indicated above, the authors of the three studies come to different conclusions regarding how to best explain the variation. We will not enter this discussion here (cf Martí 2015 for discussion of the relative merits of a "hybrid" theory vs an ambiguity theory). Rather, we will here argue that it is at best premature to conclude, on the basis of the kinds of results that we have so far seen, that a mainstream externalist, causal-historical theory of reference for natural kind terms should be replaced by a theory that posits systematic ambiguity, or context-dependence. The variation is certainly interesting, and deserves closer attention, but at the moment it is far too early to conclude anything about the fate of the causal-historical theory, and of (first-order) externalism more generally.

First of all, the studies do not look directly at the test subjects' dispositions to apply and interpret natural kind terms. Asking subjects to make metalinguistic judgements about the reference of terms, or about co-reference, or asking them to make truth value judgements, are plausibly sources of indirect data about their linguistic dispositions, but they introduce various possible sources of error (cf. Cohnitz & Haukioja 2015). We will return to this issue in the next section.

Secondly, the studies only inform us about the subjects' initial inclinations to apply (or, form metalinguistic judgements about others' application of) natural kind terms. While the variation in such judgements is, apparently, robust and in need of explanation, so far we have no reason to think of it as data to be explained, rather than as the discovery of a range of circumstances where ordinary speakers are prone to a certain kind of (fairly systematic) error. That is, we are not forced by the data to conclude that natural kind terms are, in some contexts, correctly applied as the descriptivist theory would have it. The data is equally consistent with the conclusion that, in certain contexts, speakers are inclined to classify things *incorrectly* as belonging, or not belonging, to a natural kind, or to make *mistaken* judgements about existence claims using natural kind terms, and so on.

To emphasize, we are *not* claiming that the causal-historical theorist can simply choose to ignore the variation that has been found, by dismissing the non-externalist applications as erroneous. On the contrary, we think it has to be taken seriously, and that it cries out for an explanation. But it is not obvious that the variation should be taken at face value. We know that speakers make mistakes; we know that speakers can be prone to making *systematic* mistakes in their application of language. The data that we have been presented with can not, by itself, determine whether the variation is evidence of systematic ambiguity, or of systematic speaker errors. To make progress, we need to reach a better understanding of what makes some applications of terms correct and others incorrect.

To illustrate, and to take the first steps towards developing our own view, consider a bit of anecdotal evidence. One of the authors of this paper vividly remembers his first encounter with the Twin Earth thought experiment, as a first year philosophy student. His first reaction (after a brief period of initial puzzlement) was to say that XYZ on Twin Earth is water. But very soon thereafter (possibly in response to a comment by an older student - here, recollections are unfortunately quite hazy) he realized that XYZ on Twin Earth is *not* water (given, of course, that the watery substance on Earth is H2O and not XYZ). This realization came with a clear sense of having *made an error* in the initial judgement concerning the thought experiment. That is, he wanted to take back his earlier application of 'water' to the imaginary substance on Twin Earth; he judged that, contrary to his first reaction, 'water' does not correctly apply to XYZ.

Based on our experience as teachers, and based on an informal survey of our colleagues, this experience is far from uncommon. But note that had the past time-slice of one of the authors been, prior to this realization, transported forward in time to take part in one of the empirical studies discussed above, he would very likely have contributed towards the result that natural kind terms are context-dependent, ambiguous, or governed by a hybrid theory. This suggests to us that there is a very real possibility that a number of the subjects who were used in the experiments referred to earlier were similar in this respect. Again, we are not claiming that this is the obvious explanation of the variation in judgements, and that the results can therefore be ignored. We are merely claiming that this is one possible explanation, and that the issue deserves further study.

On the dispositionalist meta-internalist view we described above, the correctness or incorrectness of particular applications of terms is determined precisely by the kinds of second-order dispositions to correct one's own past usage of a term that the above anecdote illustrates. Such dispositions can be studied experimentally, but to do that, experimental semantics would have to distance itself further

from the thought-experiment paradigm that has so far been dominant. The kinds of thought experiments that have figured prominently in the internalism/externalism debates are philosophers' tools: experiments formulated by philosophers for other philosophers who are already familiar with the competing theories and who are thereby able to focus on the relevant features of the scenarios that are (often sketchily) presented. Gathering first reactions to such scenarios from non-philosophers is unlikely to get us very far. To get more informative data, more sophisticated experimental setups should be developed, but it is not at all clear how we should go about this. In the final section, we turn to this question.

4. What would need to be done

We argued above that, in addition to a speaker's dispositions to apply terms to things in the world, three kinds of corrective dispositions are crucial for determining the meanings of her terms. To make further progress, experimental work on natural kind terms should focus on such corrective dispositions, in addition to the application dispositions studied in the experimental work referred to above.

The two first types of corrective dispositions had to do with speakers' inclinations to reconsider and revise her application of a term in response to new empirical information about the putative extension of the term, and about other speakers. As we pointed out above, *if* physical externalism and social externalism are true of natural kind terms, they are *made* true by the presence and systematicity of such dispositions in speakers using the term. Experimental work that aims to settle whether one or another form of externalism holds for natural kind terms, as used by ordinary speakers, should then focus on the relevant kinds of corrective dispositions.

First, concerning physical externalism: *are* speakers, in fact, disposed to reconsider and revise their applications of natural kind terms, should it turn out that they were wrong about the underlying features of things putatively belonging to the extension of the term? The only existing study that tries to tackle this question directly is Jylkkä et al (2009). In this study, test subjects were presented with Twin Earth style scenarios in two stages. In the first stage, they were given a story about newly discovered samples with a similar appearance to known samples of a natural kind, and told either that the new samples are found to have a similar underlying nature as the old samples, or that they are found to have a different underlying nature, despite the similarity in appearance. At this point, subjects were asked to judge whether the new samples belong to the same natural kind as the old ones, testing their dispositions to apply natural kind terms. At the second stage, the subjects were

told that the previously given information concerning the underlying nature of the old samples has turned out to be incorrect: in the scenarios where the old and new samples were first thought to differ in underlying nature, they turn out to share the same underlying nature, and vice versa. At this point, the subjects were asked whether they still agreed with their earlier classification of the new samples or not, testing the kinds of corrective dispositions we have claimed to make physical externalism true. The results gave some support for physical externalism: most subjects were disposed to revise their earlier judgement in the light of the new empirical information presented in the second stage of the scenarios. The results were, however, inconclusive, and a similar split between seemingly externalist and internalist responses was found in this study, as well.

Second, concerning social externalism: *are* speakers, in fact, disposed to reconsider and revise their applications of natural kind terms, should it turn out that the relevant experts in their linguistic community would not apply a term to things they are disposed to apply it to, or vice versa?

Third, concerning correction based on closer consideration: would subjects, for example, persist in the kinds of judgements that have been taken as evidence for systematic ambiguity, or context-dependence, were they encouraged to think more about the cases, or exposed to more cases of the same kind? Or, would they eventually converge on judgements that are more clearly in line with either the causal-historical theory, or descriptivism?

Note further that the three kinds of corrective disposition can work in tandem. If, for example, subjects who are first exposed to scenarios similar to those used by Jylkkä et al (testing the presence of the first kind of corrective dispositions), and who had responses similar to the ones reported in that study, were further to be told that the relevant experts in their linguistic community would uniformly categorize entities in one way, would they revise their application (testing the presence of the second kind)? And finally, would these judgements survive closer examination of the cases, by the subjects (testing the presence of the third kind)?

These are, then, the kinds of dispositions that *should* be looked at in order to be able to say, with any confidence, whether one or another form of externalism is true of natural kind terms, and whether there is substantial variation in natural kind concepts. So far, we have not said anything in this section about *how* such dispositions are to be studied.

First, it should be a consequence of the meta-internalist view presented that one should favour *evidence about linguistic usage* over *evidence about (tacit) metalinguistic beliefs* (Cohnitz &

Haukioja 2015). In other words, what should primarily be studied is actual production and interpretation of linguistic expressions of the relevant kind. In many studies so far, test subjects have been asked to make metalinguistic judgments of some form or other, which is at least inviting certain forms of systematic mistakes (cf. Cohnitz & Haukioja 2015).

We take it that the interpretation of linguistic expressions, as well as the production of utterances are both prone to be influenced by contextual, pragmatic factors. We might interpret others by taking their perspective, instead of interpreting them simply in accordance with what their expressions mean, and the same can, of course, happen in an experiment in which we ask the test subject to tell us how to interpret certain expressions in utterances of third parties (Sytsma & Livengood).

These potential confounds can be avoided by either moving to neutral contexts and to studies of production (this could be realized through studies of elicited production, self-paced reading, eye-tracking, etc.; for some suggestions along these lines see Cohnitz 2015). Thus, secondly, one should favour *production in contexts that are unlikely to be tinted by pragmatic factors*.

Since theories of reference concern the use of expressions already introduced into the linguistic community, and our corrective dispositions with respect to these, experiments should, thirdly, ideally *use entrenched expressions*. Most studies conducted so far instead introduce new expressions[9] (and then present conflicting information about the referents of these new expressions), but this at best tests the corrective dispositions of speakers at the point of reference fixation of a new expression, rather than how an already introduced expression refers. The worry is that with new expressions it will be hard to distinguish situations in which a speaker changes her usage because she believes that she was mistaken about its meaning and situations in which she changes her usage because she is using the expression in question with a deferential meaning.[10] However, using entrenched expressions does bring some additional complications with it. In a pilot study that lead to Jylkkä et al (2009), Jylkkä, Railo & Haukioja used a similar setup to the one used in the published work (described earlier), but with entrenched natural kind terms in some of the

9 Of the studies mentioned in the previous section, only Braisby et al (1996) uses entrenched rather than invented natural kind terms.

10 There is, of course, a general problem of distinguishing cases in which a speaker changes how she applies an expression, because she uses the expression with a deferential meaning and received new information about its proper application, and cases in which a speaker changes the meaning of one of the expressions in her repertoire, not because she used that expression with a deferential meaning but because she wants to align the way she uses the language with the wider linguistic community she belongs to. The theoretical distinction between the two situations is clear: only in the former type of situation will previous applications of the expression now be considered mistaken (such that they were mistakes all along). To tease this out empirically is not easy, but we don't see why it should be impossible.

scenarios, including a variant of Putnam's Twin Earth case. The subjects' judgements for the entrenched expressions were highly variable, and did not exhibit clear patterns. In informal post-experiment interviews with the subjects, it became clear that the subjects were very reluctant to go along with the stories and to accept them as true, when they featured familiar natural kinds such as water. The best way to make sense of a story where "scientists have found out that the clear, odourless liquid in the lakes and rivers is not H2O but XYZ" was, for many subjects, to assume that the scientists had suddenly lost their minds, rather than to bracket all their empirical knowledge about the world.[11] Needless to say, such an interpretation of the scenario made their responses worthless as data about externalism. Jylkkä et al resolved this problem by using only invented examples, but for the reasons mentioned above, it would be ideal to find other ways of circumventing this problem.

Fourth, since meta-internalism considers reference not just to supervene on usage, but on counterfactually robust usage, i.e. usage that includes our dispositions to correct what we perceive as mistakes, tests should *include possibilities for self-correction*.

5. Conclusions

In the last section we formulated what we consider to be desiderata for empirical investigations into reference. We find it difficult to even sketch concrete suggestions how these desiderata could be met by specific experimental set-ups. Reference is an abstraction from a complex set of dispositions which together constitute our use of language. Studying such an abstraction empirically is hard but we don't think that it is impossible. Although we argued above that the empirical results so far do not cast serious doubt on externalist theories of reference, we do nevertheless think that reference is an empirical phenomenon, and should be approached by empirical methods.

References

Braisby, N., Franks, B. and Hampton, J. (1996), 'Essentialism, word use, and concepts', *Cognition* 59, 247-274.

Cohnitz, D. (2015), 'The metaphilosophy of language', in: Haukioja, J. (ed.): *Advances in Experimental Philosophy of Language*, London: Bloomsbury Academic, 85-108.

11 We should add that when the reluctant subjects were explicitly told that they should accept the improbable story as true, they had no problems doing so. This reaffirms the point made earlier: the classic thought experiments were written by philosophers, for other philosophers who already know what is at stake. Studies on non-philosophers where such stories are used only with minor variations are unlikely to give us very informative data.

Cohnitz, D. and Haukioja, J. (2013), ‘Meta-externalism vs.meta-internalism in the study of reference’, *Australasian Journal of Philosophy* 91, 475-500.
Cohnitz, D. and Haukioja, J. (2015), ‘Intuitions in philosophical semantics’, *Erkenntnis* 80, 617-641.
Devitt, M. (2011), ‘Deference and the use theory’, *ProtoSociology* 27, 196-211.
Genone, J. and Lombrozo, T. (2012), ‘Concept possession, experimental semantics, and hybrid theories of reference’, Philosophical Psychology 25, 717-742.
Hampton, J.A., & Passanisi, A. (2016), ‘When intensions don't map onto extensions: Individual differences in conceptualization’, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42, 505-523.
Häggqvist, S. and Wikforss, Å. (forthcoming), ‘Natural kinds and natural kind terms: myth and reality’, *British Journal for The Philosophy of Science*.
Johnson, M. and Nado, J. E. (2014), ‘Moderate intuitionism: a metasemantic account’, in: Booth, A. R. and Rowbottom, D. (eds.): *Intuitions*, Oxford: Oxford University Press, 68-90.
Jylkkä, J., Railo, H. and Haukioja, J. (2009), ‘Psychological essentialism and semantic externalism in lay speakers’ language use’, *Philosophical Psychology* 22, 37-60.
Lewis, D. (1984), ‘Putnam’s Paradox’, *Australasian Journal of Philosophy* 62, 221-236.
Martí, G. (2015), ‘General terms, hybrid theories and ambiguity: a discussion of some experimental results’, in: Haukioja, J. (ed.): *Advances in Experimental Philosophy of Language*, London: Bloomsbury Academic, 157-172.
Nichols, S., Pinillos, A., and Mallon, R. (2015), ‘Ambiguous Reference’, Mind
Putnam, H. (1975), ‘The meaning of “meaning”’, *Philosophical Papers Vol. 2: Mind, Language, and Reality*, Cambridge: Cambridge University Press, 215-271.
Sveinsdóttir, Á. (2008), ‘Essentiality conferred’, *Philosophical Studies* 140, 135-148.
Sytsma, J. and Livengood, J. (2011), ‘A new perspective concerning experiments on semantic intuitions’, *Australasian Journal of Philosophy* 89, 315-332.
Williamson, T. (2007), *The Philosophy of Philosophy*, Oxford: Blackwell Publishing.