Experimental Semantics: The Case of Natural Kind Terms

Sören Häggqvist and Åsa Wikforss

Stockholm University

In *Experimental Philosophy of Language*, ed. J. Haukioja, London: Bloomsbury Publishing, 2015.

Introduction

Recent experimental work on judgements about reference has given rise to wide-ranging debates. But these debates have focused almost exclusively on the putative impact on theorizing about names. Natural kind terms have by and large received what might be called the Cinderella treatment. One principal aim here is to redress this neglect and see which, if any, conclusions should be drawn concerning them from experimental work so far. Another, connected, aim is to assess when and where different sorts of empirical evidence properly enter in semantic theorizing. Finally, we shall briefly consider (and lament) the status of the armchair case for current orthodoxy regarding natural kind terms.

The paper has three parts. In the first, we briefly try to sort out the general dialectic surrounding semantics and experimental philosophy established by Machery, Mallon, Nichols, and Stich (2004) – hereafter "MMNS 2004". In the second part, we attempt to disentangle certain ambiguities of the label "semantic theory" and discuss the extent to which different kinds of empirical evidence are relevant for different levels of theorizing. The final section first considers the armchair arguments motivating Kripke's rejection of descriptivism, as a semantic theory for natural kind terms; then we consider the empirical evidence concerning such terms – both from the experimental literature

and from history of science. In our estimate, available empirical evidence poses a serious challenge to Kripke's account of the semantics of natural kind terms, and lends some support to cluster theories of these terms.

1. The Origins of Recent Debates

MMNS 2004 offered tentative evidence that intuitions about reference vary with culture.[1] It suggested that this is true for intuitions of professional philosophers as well as for the participants in the study (MMNS 2004, B9). And it attempted to stir philosophers of language out of methodological complacency: instead of relying on their own, partly culture-induced and non-universal, intuitions, they should get out of the armchair. At the same time, MMNS 2004 assumed that philosophy of language relies on evidence consisting precisely of intuitions concerning reference in cases like those used in their study:

> There is widespread agreement among philosophers on the methodology for developing an adequate theory of reference. The project is to construct theories of reference that are consistent with our intuitions about the correct application of terms in fictional (and non-fictional) situations. (MMNS 2004, B3)

Now although the paper's target is obviously methodological, it is not clear that the method described above is really under attack. The authors suggest that results of the type they predicted (and found) "would raise questions about whose intuitions are going to count, putting in jeopardy philosophers' methodology" (B4). But what is jeopardized, it seems, is the assumption of universality rather than the basic method – which itself,

one might note, seems sketched almost to the point of caricature: mere *consistency* with intuitions appears a remarkably low standard for a "theory of reference".

In Mallon, Machery, Nichols, and Stich (2009) – henceforth "MMNS 2009" – the methodology to which they hold philosophers of language committed is called "the method of cases":

> *The method of cases:* The correct theory of reference for a class of terms *T* is the theory which is best supported by the intuitions competent users of *T* have about the reference of members of *T* across actual and possible cases. (MMNS 2009, 338).

Here the bar is, more plausibly, raised from mere consistency to actual support. But in this paper, the principal target are so-called "arguments from reference": attempts to "derive philosophically significant conclusions from the assumption of one or another theory of reference" (MMNS 2009, 332). Although the authors list – as one option open to philosophers in light of their results – the possibility of downplaying or giving up the method of cases, they don't advocate this but on the contrary "remain skeptical of the proposal to downplay the role of the method of cases in choosing a theory of reference" (MMNS 2009, 343). Moreover, they state that "we have no idea what other considerations philosophers of language might appeal to" (MMNS 2009, 343).[2] So if their aim all along was, as reported in Machery, Mallon, Nichols and Stich (2013, 621), to condemn the method of cases, they were somewhat coy about this. And insofar as

they wanted to condemn it while simultaneously insisting on its unavoidability, they would surely be guilty of inflicting an odd double bind on theorists.

Below, we'll try to sort out what methodological elbow room may be yet be open on various construals of "theory of reference". We shall suggest that while some construals make the theory more immediately amenable to empirical testing, along the lines of "the method of cases', all construals require a mix of empirical and theoretical considerations.

2. Experimental Semantics

2.1 The Theory of Reference

What, precisely, is "the theory of reference"? In what sense is it an empirical theory? It is important, here, to keep distinct different types of semantic theories, sometimes conflated in the literature.

The first important distinction concerns that between *descriptive semantics* and *metasemantics* (sometimes called "foundational semantics").[3] Descriptive semantics tells us something about the *semantic content* of a term: whether it is equivalent to a set of descriptions, or whether it is directly referential, etc. Semantic theories in this sense include traditional descriptivism, the cluster theory, versions of two-dimensionalism, and Millianism (among others). Metasemantic theories, by contrast, tell us something about *the facts in virtue of which* a term has a certain semantic value. Use theories of meaning, causal theories, social externalism and internalism are all metasemantic

theories in this sense. Thus, a use theory appeals to the individual speaker's dispositions to use a term, whereas social externalism emphasizes the practice of the linguistic community and causal theories the role of the physical environment in the determination of meaning and reference.

Although descriptive semantics and meta-semantics are related there is no one-one relation. For instance, even if descriptivism is typically coupled with some version of a metasemantic use theory, it is quite possible to combine a descriptivist semantics with an externalist metasemantics.[4] And while Millianism may invite some sort of causal metasemantic theory (which perhaps explains why Millianism and causal theories are so often conflated), the details of this metasemantic theory are up for grabs. Is a purely causal theory required? Or do we need a mixed version, in which speaker intentions contribute to determining reference? This illustrates that metasemantic theories are more theoretically committed than descriptive semantics, a point we shall stress below.

With this distinction in place, we can address the central question what role, if any, empirical evidence does play. Construed as a descriptive semantic theory, the theory of reference is a straightforwardly empirical theory. After all, construed that way it is a theory concerning the semantic content of a given language *L*, as used by a speaker (or a group of speakers) at a given time. To determine which such theory is true of *L* empirical evidence is required. However, even within descriptive semantics it is useful to distinguish between a less theoretically committed version and a more theoretically committed one.

The less committed theory, let us call it a "weak semantic theory", is simply one that tries to get the referents and extensions of singular and general terms in *L* straight. Such a theory is obviously closely tied to the relevant evidence (such as facts about use), but even here there is a gap between the evidence and the theory. After all, people make mistakes, and the theory has to take this into account. Therefore, some apparent evidence provided by use may have to be written off. The task of the philosopher, then, is to provide some principle that serves to map facts about use on to meanings, allowing us to determine what counts as an error and what does not.[5] Even when it comes to weak semantic theories, therefore, there are important theoretical commitments involved. Nevertheless, the theory as a whole must be based on empirical facts about the speaker use.

A more committed version of descriptive semantics – "strong semantic theory" – goes further, and assigns semantic content in accordance with theories such as the Millian theory or some version of descriptivism. This is one more step removed from the evidence, since it is possible for both Millians and descriptivists to accept the same weak semantics. That is, they may agree on the truth conditions of sentences as used by a speaker (even "across possible worlds") but disagree on what semantic theory best accounts for these truth conditions. Indeed, the versions of descriptivism that emerged after Kripke typically tried to accommodate his claims about the reference of proper names, for instance by appealing to rigidified descriptions.[6] It should be clear that this is a kind of disagreement that cannot be settled by further evidence from speaker use. Other types of considerations have to be marshalled here, such as various theories' explanatory power visavi phenomena like co-referring or empty names. Arguably, other

sorts of empirical evidence are also relevant, such as findings within developmental psychology or cognitive science.

Next, moving on to metasemantics, we are one step further removed from the empirical evidence. Since the correct semantic theory (weak and strong) does not by itself dictate a certain metasemantic theory, general epistemological and metaphysical considerations will play an important role here. One example is the debate concerning self-knowledge and semantic externalism, where many philosophers accept that if they are incompatible, we should question the externalist metasemantics. Nevertheless, metasemantics too must be considered an empirical theory, constrained by speakers' application of terms, both to actual and hypothetical cases (such as Twin Earth cases). So while metasemantic theories are less directly connected to the empirical evidence, and more dependent on general philosophical considerations, they are clearly empirical theories, subject to falsification. [7]

How do MMNS construe "theory of reference"? Are they testing weak semantic theories, strong semantic theories or metasemantic theories? This, unfortunately, is not clear. The talk about descriptivism and the focus on Kripke's Gödel-case suggests that their concern is with weak and strong semantic theories. After all, Kripke appeals to intuitions concerning the counterfactual story about Gödel, in order to show that proper names are rigid and to provide evidence against descriptivism in support of Millianism. At the same time, MMNS claim to be testing the *causal-historical view*, not Millianism (2004, B4-B5); and this, as we saw, is a metasemantic theory.[8] However, as stressed above, there is only a loose connection between Gödel-style intuitions and the causal-

historical theory. Before a purely causal theory could be said to be supported, for instance, it would first have to be shown that the semantic theory that best accounts for the Gödel-intuition is Millianism, rather than some version of rigidified descriptivism or two-dimensionalism. Then it would have to be shown that the metasemantic theory that best accounts for Millianism is a pure causal-historical theory, rather than some mixed theory (appealing to speaker intentions in addition to causal links). The connection between the empirical evidence and the theory of reference, understood as a metasemantic theory, is therefore less close than MMNS (2004) and MMNS (2009) appear to hold.

This unclarity underlies some of the current disputes over their experiments. Thus, Martí (2009) has argued that the experiments carried out are flawed and do not prove what they purport to prove, since they test subjects' meta-linguistic judgements.[9] Martí is surely right to draw attention to the fact that MMNS are asking speakers to make meta-linguistic judgments concerning what they think a term refers to. If the aim is to determine the correct semantic theory for a class of terms, it would seem more relevant to consider how speakers in fact apply these terms. But, as noted above, the aim of MMNS (2004) was to examine judgements supposed to be indicative of people's theories of how their terms refer, their "folk-psychological metasemantics". To this extent, Martí's objection misfires, as Devitt (2011, 428, fn. 8) notes. At the same time, there is something strange about the project understood this way. Why should we expect ordinary people to have views, even implicit ones, on *how their terms refer*? The parallel with Chomskian theory tentatively broached by MMNS (2004, 88) does not hold up: even if it is accepted that ordinary speakers have an implicit semantic theory,

guiding their use of language, it is quite different to claim that ordinary speakers have an implicit *metasemantic* theory.

If MMNS really aimed to test *semantic* theories (as they sometimes seem to claim), then Martí's objection would be justified: to test semantic theories (weak and strong), what would seem to be needed is the whole artillery of tests that linguists employ, including corpus studies and elicited production. As for judgements, we should rely simply on first-order judgments involving the application of proper names to individuals and the use of general terms in categorization tasks, not reflexive judgments *about* the use of terms.[10]

Another objection to MMNS (2004) can be found in Deutsch (2009). According to Deutsch, the very idea that intuitions play an important role in semantics is simply a mistake: What matters is philosophical *theorizing*, not intuitions.[11] Kripke's arguments against descriptivism, Deutsch argues, simply rely on counterexamples, not on intuitions. For instance, that "Gödel" refers to Gödel, not to Schmidt, is a counterexample to this theory since descriptivism predicts that "Gödel" applies to Schmidt (in these circumstances). But when Deutsch suggests that the philosopher can take for granted facts about reference, he is in effect suggesting that she needn't worry at all about the first step, that of determining the weak semantic theory. The problem here is that even if some meta-linguistic claims may seem too trivial to require any empirical evidence – such as the statement that "Gödel" refers to Gödel – they are nevertheless empirical claims and as such they do, ultimately, depend on empirical evidence.[12] And when we move to general terms, including natural kind terms, this step

is more controversial as well as more theoretically involved. For example, the claim that "water' only has H_2O in its extension cannot be presented as a counterexample to descriptivism unless we have already been given some reasons to think that "water" only has H_2O in its extension. This is far from trivial.

2.2 The Expertise Defence

Much of the evidence adduced by experimental philosophers (including MMNS 2004), concerns lay persons, not philosophers. Hence the claim that it threatens traditional philosophical methodology is vulnerable to the objection that responses, intuitions, or opinions of lay people are not the relevant evidence in the first place. This objection, made by several writers, often uses analogies with other disciplines. For instance, Devitt writes: "We don't do physics, biology, or economics simply by consulting people's intuitions. Why should semantics be different?" (Devitt 2011, p. 424).[13]

We have stressed that doing semantics is never a matter simply of constructing a theory "consistent with our intuitions about the correct application of terms", as MMNS (2004, B3) suggest. Even when it comes to weak semantics, theoretical virtues and general philosophical considerations must be brought to bear. This arguably takes expertise. However, it is one thing to recognize this point, quite another to suggest that ordinary intuitions are irrelevant or less weighty than expert intuitions. Since semantic theories are *empirical* theories, contingently true of a group of speakers at a time, there is simply no possibility of doing semantics without evidence provided by speaker use. Whatever

the merits of the expertise defence in other areas, it is of limited value when it comes to semantics.

It is sometimes suggested that the trouble with the expertise defence in the case of semantics concerns an assumption that expert intuitions are more reliable because they are the product of better theories. *Linguistic* intuitions, it is argued, reflect the linguistic competence of the speaker and it makes no sense to assume that the philosopher has a greater linguistic competence than the ordinary speaker (MMNS 2013, 627). However, we do not think it is very helpful to speak of linguistic intuitions as merely reflecting the linguistic competence of the speaker. How we apply our terms is of course in part a reflection of our linguistic competence but equally a reflection of our general knowledge and background beliefs, and it does not seem possible to peel off a part of this use as being an expression specifically of the speaker's "linguistic competence". For this reason, it is quite possible that the better one's theory in a particular area, the more reliable are one's "intuitions" about how the relevant term applies. Thus, the philosophers have a greater competence when it comes to philosophical terminology, just as the biologists do when it comes to biological terminology; for instance, in order to understand the semantics of "tiger", as used in the common language, it makes sense to pay attention to how biologists use this term in biological classification.

This is not to deny that there may be relevant differences in the lay use of kind terms and the expert use of these terms. However, the very point of the Kripke-Putnam account of natural kind terms is precisely that our everyday use of these terms is *continuous* with their use within science, and that what the terms ultimately "pick out",

supposedly determined by the essence of the kind, is a matter for science to adjudicate.[14] If so, the intuitions that carry weight are those of the relevant scientists, not of the philosophers.

It might be suggested that even if this is so, there is reason, when it comes to certain types of semantic intuitions, to trust the philosophers more than the layperson. Devitt (2011), for example, suggests that when it comes to *modal* intuitions, these are likely to be more reliable if the subject has reflected on metaphysical matters, such as the essence of individuals or the essence of natural kinds, and so "the intuitions we need are ones from people with some expertise in these matters, presumably metaphysicians and other philosophers" (Devitt 2011, 427). Since arguments for or against Millianism depend crucially on evidence provided by intuitions about what is possible, what is necessary, and what would be the case in various hypothetical scenarios, expert intuitions will play a decisive role, if Devitt is right.

Now, it is clear that it is an open question to what extent the modal intuitions of non-experts are reliable. But the fact that there is a great deal of variation in important modal intuitions among philosophers suggests that these intuitions are not entirely reliable when held by experts either (assuming the language is shared).[15] Moreover, to the extent that Devitt is right and we need to have done metaphysics in order to have modal intuitions, the value of these intuitions as evidence for the semantic theory decreases. For example, if my modal intuitions about gold or water require a theory concerning the essences of natural kinds, then these intuitions run a real risk of being theory driven, as

Devitt explicitly thinks they are. If so, we should handle modal intuitions with care rather than accepting them as theory independent evidence for semantics.[16]

Again, none of this is to deny general philosophical considerations matter when doing semantics. For the same reason, the mere fact that there is variation in speaker judgments does not immediately imply that there is a semantic variation. Philosophical theory plays a central role when constructing the semantic theory (even a weak one) out of the materials provided by speaker use. What is wrong with the expertise defence, it seems to us, is not the idea that expertise matters, but the idea that when it comes to basic applications of terms, the philosopher's intuitions carry greater evidential weight than those of the non-experts. We now turn to the case of natural kind terms, where it has been widely assumed that we can do semantics from the armchair.

3. Natural Kind Terms and Empirical Evidence

3.1 Kripke's Armchair Intuitions

Kripke's main target in the case of proper names was the cluster theory. He does not merely argue against definitionalist versions of descriptivism, but follows Mill and argues against the very idea that proper names have descriptive content. Unlike Mill, he applies the same strategy in the case of natural kind terms, arguing against the cluster theory and in support of a Millian semantics.[17] In this, Kripke departs radically from his contemporaries. After Quine, there was widespread skepticism about traditional, definitionalist accounts of general terms and versions of the cluster theory were suggested to take its place.[18] Kripke breaks with all this, arguing that natural kind terms are semantically akin to proper names and lack all descriptive content.

One would expect a radical claim of this sort to require substantial support. However, Kripke's defence of the Millian account of natural kind terms is sketchy and brief (covering barely 25 pages of the book). It raises a number of questions. In particular, why should natural kind terms, being general terms, be treated like proper names? And what does it mean for a kind term to be a rigid designator? The latter question has been much debated since Soames (2002), but there is little consensus on how it is to be answered. We shall leave it here and simply consider the underlying question: Why should natural kind terms be given the same Millian semantics as proper names (assuming that is the correct semantics for the latter terms)? What is Kripke's evidence for taking this radical step?

Kripke starts the discussion of natural kind terms by criticizing Kant's claim that the judgment "Gold is a yellow metal" is analytic and a priori. Kripke imagines a scenario in which it is discovered that the yellow appearance of gold is merely an illusion, and argues that if this were to happen we would conclude not that gold does not exists, but simply that it has turned out that gold is not yellow. This is an empirical claim about how we would respond, but Kripke's armchair guess seems prima facie plausible. After all, we have accepted that some gold is white. But this type of scenario merely provides evidence against *definitionalist* versions of descriptivism, not against the cluster theory. Kripke is aware of this and quickly moves on to make two much more radical claims:

> (i) Something could have *all* the properties normally associated with gold and not be gold.

(ii) Something may have *none* of the properties associated with gold and still belong to the kind.

It is clear that (i) and (ii) are needed to support Millianism since the central Millian contention is that natural kind terms lack all descriptive content: whatever descriptions we associate with these terms, their function is merely to *fix* reference (i.e. the metasemantic function of fixing the term onto a property in the actual world), not to *determine* reference (i.e. the semantic function of determining reference and extension across all possible worlds). Instead, the "underlying" property itself (however that is to be understood) somehow provides necessary and sufficient conditions, whether or not speakers have any knowledge of this property.

Thesis (i) concerns the possibility of "twin cases", of the sort discussed by Putnam. To support (i) Kripke suggests that if there were a substance which had all the identifying marks of gold, but which is not the same substance, we would say that "though it has all the appearances we initially used to identify gold, it is not gold" (1980, 119). This seems true, but trivially so. If something is a *different substance*, and we are aware of this fact, we would not call it "gold". Less trivially, Kripke appeals to a real life example, that of fool's gold which, he claims, has all the appearances we initially used to identify gold (1980: 119, 124). The trouble with this example, as many have pointed out, is that fool's gold does *not* have all the appearances of gold. On the contrary, the substances have been seen as distinct long before the development of modern chemistry and knowledge of atomic numbers.[19] Consequently, even if widely shared, the intuition that fool's gold is not gold does not support the claim that something may have *all* the

normal properties of gold and not be gold. Indeed, any difference in underlying properties appears bound to cause differences in macro-level properties as well.[20] It is therefore no accident that philosophers have taken recourse to thought experiments to test thesis (i).

Like Putnam, Kripke proposes such a thought experiment. He imagines a tiger that has all the characteristics typically associated with tigers, all the external appearances of a tiger, but with an internal structure completely different from that of the tiger, the structure typical of reptiles. Raising the question whether we would conclude that some tigers are reptiles, he responds: "We don't. We would rather conclude that these animals, though they have the external marks by which we originally identified tigers, are not in fact tigers, because they are not of the same species which we called 'the species of tigers'" (Kripke 1980, 120). This claim about what we would conclude in such a scenario, clearly, is a strong empirical claim. If indeed the aim is to determine the semantics of natural kind terms, as used by ordinary speakers, the claim would have to be tested: Kripke's armchair judgments cannot by themselves be decisive.[21]

Moreover, Kripke's judgments are in conflict with what the experts – i.e., biologists – say about the individuation of species. Kripke associates forming a natural kind with having a particular internal structure. But the notion of internal structure makes little sense in the case of species, and however it is construed, few biologists would take it as criterial.[22] Kripke therefore has to choose: Either he can insist that species terms track "internal structure' (in accordance with his intuitions) or he can hold on to the idea that the extension of natural kind terms is ultimately determined by science.

Similar worries apply to thesis (ii), which is by far the stronger thesis. In the case of proper names, Kripke appeals to modal intuitions, such as the intuition that Aristotle might not have had *any* of the properties commonly attributed to him (Kripke 1980, 75). In the case of natural kind terms, similarly, Kripke appeals to the intuition that tigers might have had none of the properties by which we originally identified them. It is possible, he says, that tigers lacked all the properties we commonly attribute to them, such as being tawny yellow, striped, carnivorous, fourlegged, and so on. For instance, Kripke suggests, we might find out that all of these properties had been mistakenly attributed to tigers due to "optical illusions or other errors". From this Kripke concludes that "gold" and "tiger" do not mark out a cluster concept in which most of the properties used to identify the kind must be satisfied: "On the contrary, possession of most of these properties need not be a necessary condition for membership in the kind, nor need it be a sufficient condition" (Kripke 1980, 121).

In support of thesis (ii), therefore, Kripke simply appeals to his intuition that tigers could lack all of the properties normally attributed to them, without spelling out any details. Is this a widely shared intuition?[23] Do we even know what we are supposed to imagine – a creature that does not have four legs, is not tawny yellow and striped, does not have large teeth, is not a carnivore, and is not a mammal? One without fur, skin and eyes? And how does this proposal square with the classifications employed by biologists? Kripke, no doubt, is driven to this conclusion by his assumption that natural kind terms are semantically more similar to proper names than to other kind terms. But, again, what is needed is precisely some empirical evidence, or at least argument, to

support this claim, which for over four decades has shaped much of what philosophers of language have said about natural kind terms.

Kripke's assimilation of natural kind terms to names has been momentous, despite his caution when suggesting it (1980, 127-8). On the one hand, it has led to attempts to construe natural kind terms simply as singular (for a recent attempt, see Bird 2009; for a rebuttal, see Needham 2012) or at least as rigid (for rebuttals, see Soames 2002). On the other, it has more insidiously invited the thought that *something* must confer on kinds the sort of metaphysical unity across possible worlds intuitively had by individuals. Many have found it tempting to think of microstrucure as offering such unity; hence the allure of (micro)essentialism. Kripke's discussion certainly seems informed by something like this idea and as we shall see shortly, it is easily made even when it shouldn't. We think that the semantic thesis underlying it – that natural kind terms function in a way completely different from other general terms – is mistaken.

We have not ourselves carried out any experiments, so we don't have any new data to offer here. Instead, we shall survey the available experimental data and other evidence. The extant data suggest that Kripke's picture of how natural kind terms function does not cohere with ordinary use of these terms (within or outside science) and that a much more complex picture is required, one that does not fit with the Millian proposal.

3.2 Empirical Evidence from Experiments

Experimental psychologists have been interested in natural kind concepts since the 1970's, in part as a result of the work of Kripke and Putnam. The primary concern of

the psychological literature, naturally, is not semantics but "concepts", the psychological states and processes that guide people's categorization judgments.[24] A main goal has been to test *psychological essentialism*, the thesis that people take objects to have deep-lying, "invisible" essences determining categorization.[25] Early experiments provided initial support for psychological essentialism (Keil 1989, Rips 1989, Gelman & Wellman 1991), suggesting that people categorize not only on the basis of similarity but also on the basis of assumed deep-lying causal features, such as chromosomes. However, later experiments have shown the results to be less robust and actually highly contextual (Kalish 1995; Hampton 1995; Hampton, Estes and Simmons 2007).

The relation between psychological essentialism and Kripke's semantics is not straightforward, since the former is a psychological thesis, not a semantic one. Nevertheless, the experiments carried out by psychologists are relevant to semantics, since they focus on categorization tasks and typically involve term application. Many involve "name appropriate judgments", where participants face a "forced choice" of applying or withholding a term in accordance with one feature rather than another (for instance, in accordance with deep-lying features or observable features); other experiments involve "free naming", where people apply a term in spontaneous discourse.

Of particular interest, from our point of view, are psychological experiments explicitly designed to test the essentialist approach to the meaning of natural kind terms. Thus, Braisby, Franks and Hampton (1996) set out to "evaluate the intuitions about word use that Kripke and Putnam deploy in support of essentialism, by putting them to empirical

test" (1996, 248). "Essentialism" is here construed as the claim that essential properties determine reference independently of people's beliefs, and the experiments test the extent to which people allow underlying features to govern their use of natural kind terms such as "cat", "water', "tiger" and "gold". The experimental results, Braisby et al. argue, do not support essentialism of the Kripke-Putnam sort. Although in some scenarios the majority (73%) took underlying features to play an important role, a substantial minority did not, and the experiments do not support the conclusion that *only* underlying features play a decisive role (as Millianism suggests).[26]

Clearly, results of this sort are relevant when we try to determine the proper semantics of natural kind terms, and it is surprising that philosophers have largely ignored the findings of experimental psychologists. If a substantial group of people do not apply their terms in the ways predicted by Kripke and Putnam, then this poses a difficulty for their semantics. Naturally, it might be possible to explain away the divergencies – either as resulting from error or as indicating that people use terms with different meanings. However, there is one type of explanation that does not seem very promising in the context: the claim that people are mistaken in their modal intuitions, since they are not philosophers and therefore ignorant about essences of natural kinds. As we've stressed, philosophers are not the experts when it comes to natural kinds: scientists are. And even if a particular philosophical theory about natural kind essences were accepted, it would not tell us anything about the *semantics* of terms such as "water", "gold" and "tiger". Rather, if people do not use these terms to track "underlying" properties only, then it follows that they do not use them to track natural kinds of the sort envisaged by Kripke or Putnam.[27] [28]

In experimental philosophy, so far only a small fraction of studies have focused on natural kind terms, compared to names. Here we shall briefly summarize the findings of two interesting studies – Jylkkä, Railo, and Haukioja (2009), from now on referred to as "JRH"; and Genone and Lombrozo (2012) – and their implications.

JRH set out to test what they label "externalistic essentialism". This is the thesis of psychological essentialism, combined with the semantic claim that "the speakers take possessing E [the essence] to be a necessary and sufficient condition for belonging in the extension of C" (2009: 41). Their starting point is the experiments reported in Braisby, Franks, and Hampton (1996), and they argue that the study's evidence against externalism is weaker than the authors suggest and even lends some support for externalism. Moreover, JRH suggest, the internalistic answers could be explained away, perhaps as a result of the fact that the participants relied on pre-discovery identificatory knowledge associated with the kind (i.e. on the descriptions which serve to fix the extension of the term).

JRH therefore proceed to design their own experiments, hoping to secure more decisive evidence for or against externalist essentialism. The first experiment has two different stages, using a set of scenarios involving six fictive natural kinds. In the first stage, subjects read descriptions of a natural kind having certain macro-level properties, believed by scientists to be a certain compound (2009, 49). Thus in one scenario the subjects read about a yellowish, bitter-smelling, fragile mineral common in Siberia, "zircaum", which scientists believe to be the compound ACB. They are then told that in

Norway a deposit of a mineral with the same macro-level features is found, and after examining its deep structure scientists conclude that it is ACB too. Participants are asked whether they considered the novel sample as falling under "zircaum" as well, and by and large people answered yes to this question. At the second stage subjects are presented with a scenario where it turns out that the scientists had been wrong about the mineral in Norway and that it is in fact a different substance, KML. Participants are then asked to answer two different questions: Whether they consider the earlier categorization judgment to be *justified*, and whether they consider it to be *correct*. According to JRH (2009), strict externalistic essentialism predicts that while subjects may consider the judgment to be justified, they will not consider it correct, whereas internalism predicts that the participant considers her earlier judgment unambiguously correct (since the Norwegian mineral fits all the descriptions concerning the macro-level properties of zircaum).[29] The authors suggest that externalism was largely confirmed: 69% of the answers were externalistic (out of which 33% gave purely externalistic answers), 28% were internalistic, and 3% were compromises (JRH 2009, 52).

JRH (2009) comment on the need to explain the substantial number of internalist answers. One hypothesis is that natural kind terms are ambiguous and have two senses – one internalistic, another externalistic – as suggested by hybrid theories. The authors therefore designed a second experiment, intended to test the hybrid theory more directly, also reported in JRH (2009). The scenarios were similar but participants were given the option of giving an ambiguous answer ("yes on the one hand, but…"), which is what a hybrid theory would predict. Again, participants were asked whether the earlier categorization judgment was correct and whether it was justified. In this experiment too,

the results were rather mixed: 48% externalistic answers, 22% internalistic, 17% ambiguous ("yes on the one hand…'), and 12 % were "cannot say"-answers. Since ambiguous answers and internalistic answers contradict one another, the latter cannot be explained away as an expression of natural kind terms having two senses, an internalistic and an externalistic. A quarter of the answers, therefore, support pure internalism. Nevertheless, JRH (2009) conclude that since the majority of the judgments were in accordance with the predictions of externalistic essentialism this experiment also supports externalism and that, therefore, the two experiments support the claim that "category membership is determined by the hidden, external essence or deep structure *even if nobody knows about it*" (JRH 2009, p. 58).

In conducting these experiments Jylkkä, Railo, and Haukioja have provided a very important contribution to semantics. They have taken seriously the fact that externalist essentialism is an empirical thesis and set out to test it. However, we think that there is some reason to question their conclusion that the experiments provide support for externalist essentialism.

First, there is a problem with how the scenarios are described. Consider the following description of what happens when the scientists examine the Siberian substance more closely:

> Using methods and instruments more exact than previously available, they find out that they were wrong about the deep structure of the substance: the substance

> is KML instead of ACB. However, the substance found in Northern Norway was indeed ACB, just as the scientists thought it was (2009, p. 49).

By talking about the substance *being* KML, rather than ACB, it is implied that chemical substances are *identical* to their molecular composition. However, if so, it would seem to follow rather trivially that the substances are distinct: if the first substance is identical to ACB then it must be distinct from the substance which is identical to KML (assuming, as stipulated, that ACB≠KML). Subjects therefore appear *primed* to conclude that the Norwegian mineral samples do not belong to the extension of "zircaum". Hence, the description of the scenario is problematic in just the same way that Kripke's description of twin-gold is, when he describes it as a substance which has all the appearances of gold but which is not the same substance.

Moreover, there are problems with the interpretation of the results of the experiments. The first, and most obvious, difficulty concerns the fact that even if the majority of the answers comply with the predictions of externalism, a substantial minority complies with internalism, and as the second experiment shows, this cannot be explained away by appealing to a hybrid theory. In the end, the authors appear to fall back on a version of the expertise defense: "Lay speakers' intuitions towards or against externalism may be very implicit, and fuzzy, which could result in significant deviation in the answers. In contrast, philosophers have spent lots of time reflecting their intuitions and formed a very explicit opinion about the subject matter" (JRH 2009, 58). This suggests that what is decisive is not so much how non-experts judge these cases but how philosophers judge them, and that since the majority of philosophers are externalists that provides

evidence in support of externalism. This would be a step back from the empirical approach motivating the authors, and since the terms here are natural kind terms, there is – as we noted – no special reason to think that philosophers are particularly adept at using these. Indeed, many claims made by philosophers about natural kinds are simply incorrect (such as the claim that biological species are individuated by genetic structure) or disputed by philosophers of science (like the claim that chemical substances are individuated by molecular composition).

The second worry concerns the description of internalism. It is said that internalism predicts that underlying composition does not matter to categorization and that "internalism has no way of accounting for any significant number of externalist answers" (JRH 2009, 48), but this is incorrect. Consider the zircaum scenarios again. The subjects have been told that the scenarios involve substances, natural kinds, with a certain chemical composition. It has also been made clear that these kinds are important to scientists (the chemical composition is examined by them) and, therefore, that they belong to scientific categorizations and not just to everyday categorizations. Given all this, one can account for the participants' reactions simply by appealing to a version of the cluster theory. After all, if "kind K is a substance", or "kind K is a natural kind" belongs to the cluster of descriptions, and the subjects are assumed to believe that substances are not *merely* individuated in terms of their macro-level properties, then the cluster theory predicts that if the mineral in Norway has a distinct composition, a number of subjects will judge the mineral not to be zircaum. This is so even if the subject does not have any belief about the precise nature of the chemical composition.

Indeed, it seems to us, the cluster theory gives a *better* account of the mixed results of the experiments than either externalism or the hybrid theory (which, again, predicts ambiguity). According to the theory a sample falls in the extension of a kind term if it fits the weighted majority of the associated properties (this is how Kripke characterizes the cluster theory), which implies that both macro-level properties and underlying features, such as chemical composition, play a role. The precise role played by these different components, however, will depend on how individual speakers *weight* them, and it is likely that some speakers will consider the underlying component to be more decisive than others – hence, the theory predicts that some people will judge the Norwegian samples not to belong to the kind zircaum whereas others will judge them to belong to this kind.

The results reported in JRH (2009) thus illustrate an important point pressed earlier. To refute the cluster theory it is not sufficient to show that "underlying" properties *matter*: it also has to be shown that *only* underlying properties matter, as suggested by Millianism. This, however, isn't shown by the experiments reported in JRH (2009). What is being tested is only the first part of the Millian theory, thesis (i); i.e. the thesis that two samples may share all the same observable properties and yet not belong to the same kind as a result of a difference in underlying properties. Even if we ignore the above objections to the experiments, therefore, they simply do not test externalistic essentialism, i.e. the thesis that "the speakers take possessing E [the essence] to be a necessary and sufficient condition for belonging in the extension of C" (2009: 41). To test this thesis, one has to determine whether subjects would grant that a sample may

belong to a category even if not only one (or some) of the ordinary descriptions fail to hold of the sample but they *all* fail, as in Kripke's reptile-tiger case.

A similar reading is invited by the second study, Genone and Lombrozo (2012). They conducted two experiments; both concerned with kind terms for fictive natural kinds (diseases, minerals) and nominal kinds (artifacts, legal documents). In this study, undergraduates were asked to judge whether two fictional persons in various scenarios are thinking about the same disease/mineral/artifact/legal document (2012, 724-725).[30] In the first experiments, participants were asked to judge simply whether or not reference was shared; in the second, they were asked to indicate degree of agreement with a claim about shared reference on a 7-point scale (2012, 728-729).

For each term, four vignettes were used, in which the characters had variously matching – and variously erroneous – descriptive information associated with the term and where they were either linked to a common causal chain or not. The central finding, replicated in the second experiment, was that speakers do not consistently rely on either causal or descriptive information.[31] Moreover, experiment 1 found that willingness to attribute shared reference decreased as a function of number of false beliefs on one character's part; experiment 2 confirmed this finding (2012, 730).[32] If Kripkean Millianism were correct for kind terms, there should be no reason to expect this. For according to Millianism, members of a kind may lack *all* properties speakers believe that they have. Certainly, the properties that were varied in these experiments – conductivity, shininess, hardness and color – are exactly the sort that Kripke held speakers may be massively in

error about *in toto*. Hence, Genone and Lombrozo may also be read as providing tentative evidence against Millian kind term semantics.

Genone and Lombrozo suggest that individual variation may be explained by different individual strategies for weighting or combining causal and descriptive information in reaching reference judgements, and even applying different strategies on different occasions due to contextual factors. On this suggestion, statistical differences between cultures might be explained by differing, and partly culturally induced, ways of filling in unspecified contextual assumptions about presented scenarios (2012, 732-733). We think that, albeit speculative, this hypothesis is promising (perhaps more so than the speculation, inspired by Nesbitt, about wholesale cultural differences in causal thinking broached in MMNS 2004)**.** We note that it is quite compatible with a cluster theory for kind terms.

3.3 Evidence from History of Science

The broader empirical evidence for evaluating the semantics of natural kind terms is not limited to surveys using cases. Other relevant evidence includes history of science. The relevant literature here is too vast to summarize or even list, of course, but many particularly interesting data are gathered and discussed in LaPorte (2004). LaPorte writes:

> Putnam and Kripke prompt intuitions about the proper use of a term in counterfactual scenarios by asking *what we would say* were we presented with this

> or that scenario. This procedure seems reasonable. Speakers' dispositions indicate the proper use of a term. (LaPorte 2004, p. 97; italics in original)

Now irrespective of the propriety of the method of cases in general, it seems right to impute use of it to Putnam and Kripke, as MMNS (2004) did and as LaPorte does here. Invoking, as Putnam (1975) did, the parallels between H_2O and XYZ, on the one hand, and jadeite and nephrite, on the other, LaPorte immediately continues:

> And Chinese speakers have indicated by their actions *what they would say* were they presented with a new substance like what they had called "jade" except in its microstructure. These speakers have been presented with such a substance, and they have displayed a strong disposition to count it "jade". Therefore, speakers' dispositions can hardly be said to make a case for the position that [the newly discovered substance of] jadeite would clearly have failed to belong in the extension of "jade". (LaPorte 2004, pp. 97-98)

LaPorte's claim, which we endorse, is that this carries evidential weight for claims about what speakers would be disposed to say, were they presented with XYZ. It suggests that they would not go with microstructure, as assumed by Putnam (1975).[33] In fact the convoluted history of "jade" also suggests that speakers would not go with observable characteristics. The historical evidence suggests that neither counterfactual about "what we would say" is warranted, and indeed, that both are false. What the case – and others studied by LaPorte – indicates is something rather different: that both microstructural and observable properties are important, though neither is decisive, and

that when they are discovered to come apart, resulting usage may follow either and hence will be hard to predict. This is just what would be expected on a cluster semantics for kind terms, when microproperties are allowed to partake in the cluster of descriptions.

LaPorte is chiefly interested in biological kind terms. Consider terms for species, or higher taxa such as "mammal". If "mammal" has only "live-bearing" as its sole, criterial description, inclusion of egg-laying monotremes will indeed have drastic consequences (LaPorte 2004, 114-116). Insisting on applying such a single criterion would be to overzealously apply a crude definitionalist descriptivism. But happily, it is not what scientists do, or have done. As LaPorte notes, cluster descriptivist theories allow "some needed conceptual continuity" (2004, 117). They also allow some referential continuity or stability, unlike simple description theories. Key to seeing this is accepting some negotiability – and vagueness – in what the extension of a kind term includes. When Putnam, and other causal theorists, discuss reference of natural kind terms by first using the term or the generic definite description "the kind" and then moving on to anaphora like "it", they in effect exclude the possibility of open texture.[34]

LaPorte argues forcefully that the causal theory is, "contrary to wide acclaim, useless in blocking instability" (2004, 118). This is so because even if terms are introduced via baptisms of samples supposedly fixing reference conditions, *and* samples of enough foils to deflect some of the so-called "*qua* problem" (avoiding reference to metals or elements in general, rather than to gold; or to mammals or vertebrates rather than to horses), the speakers performing the baptisms cannot foresee how new discoveries will

force refinements in usage affecting the extensions of the terms. As LaPorte stresses, open texture cannot be avoided even if it is agreed that kinds have essences and what, in general, constitutes the essence of (say) species and substances. The refinement of the extension of "mammal" that resulted upon the discovery of monotremes, for instance, was not dictated by acceptance of cladistic essences for biological taxa (2004, 119). Neither was the refinement in how the term "water" came to be used after the discovery of deuterium and heavy water. Such developments always contain an element of stipulation, and so are, in effect, decisions rather than discoveries – pace Putnam and Kripke.

LaPorte attributes the causal theory's inability to guarantee referential stability to the fact that, on this theory, "baptisms … are performed by speakers whose conceptual development is not yet sophisticated enough to allow the speakers to coin a term in such a way as to preclude the possibility of open texture, or vague application not yet recognized" (2004, 118). But this observation, while sensible, overlooks a more general point, viz. that no amount of conceptual sophistication will endow speakers with an ability to – somehow – anticipate and resolve in advance potential vagueness laid bare only with future empirical discoveries. Open texture appears to be endemic. Further illustration is offered by the historical vicissitudes of virtually any putative natural kind term.

Take "acid", whose trajectory from Boyle in the 17th century until the mid-50's is traced in Stanford and Kitcher (2000, 115-118), from whom we'll borrow in this paragraph. Boyle characterized acids via descriptions of observable properties: they are sour,

corrosive, and precipitate sulphur from sulphide solutions. Subsequent refinements were principally due to, in turn, Arrhenius in the 1880's, Brønsted and Lowry in the 1920's (Brønsted 1928), and G.N. Lewis in the 20's and 30's (Lewis 1938). Lewis characterized acids as electron pair receptors, dropping earlier theorists' restriction to hydrogen compounds. His chief motivation was that this characterization captures four phenomenological criteria: (i) rapid combination with bases, (ii) replacement of weaker acids, (iii) characteristic effects on coloured indicators [like Litmus paper], and (iv) characteristic catalytic action. This rationale was defended by Luder and Zuffanti (1946, 3, quoted in Stanford and Kitcher 2000, 117): "a substance that exhibits the properties of an acid should be called an acid, regardless of preconceived notions about the dependence of acidic properties on a particular element". Lewis himself appears to have regarded the micro-level characterization in terms of electron pairs as incidental, albeit handy: "it is possible . . . to discuss and define acids and bases merely from their behavior in chemical reactions without any theory of molecular structure" (Lewis 1938, 293, quoted in Stanford and Kitcher 2000, 116).

This offers a good illustration of what happens with a term as scientific knowledge matures. A kind term is first associated with descriptions, and if the term is introduced at a stage of relative scientific underdevelopment, these descriptions typically concern observable properties. What happens with accumulated discoveries and increasing scientific sophistication, however, is *not* that microstructural characterizations are first found and then explicitly appointed to the office of determining reference or extension. What happens is rather that simple, observational descriptions are replaced by increasingly sophisticated, theoretical ones, and that microstructural descriptions enter if

they are deemed economical explanantia of other properties. Microstructure[35] may be made "criterial", but that status depends on which other properties are held to need explanation and are retained (like Lewis's four phenomenological properties) and which are deemed expendible (like Boyle's sour taste), and this – as LaPorte stresses – is highly contingent and involves some negotiation and decision-making. It is not a matter of simply "discovering that some members of a natural kind lack properties originally used in picking out a kind", as Stanford and Kitcher suggest (2000, 117). Of course, if a fairly stable characterization of a kind is in place, and "pick out" just means that some property is thought to be a reliable indicator of a kind, is seems fair to call this simple discovery of certain kind members' lacking this property – say, that not swans are white or not all zebras striped. But dropping the hydrogen requirement, in the case of "acid", was not a matter of simple discovery – to judge from the records, the transition from Brønsted-Lowry to Lewis didn't even have a phenomenology of simple discovery to the chemists involved.

Moreover, microstructural criteria may not emerge for natural kind terms, either because no candidate is found or because suggested candidates are contested, due to disagreement over the importance of other descriptions.[36] The history of science is, as we said, replete with illustrations of both sorts of cases.[37] But on the other hand, this matters a great deal less for science than would often seem assumed by philosophers of language. Firstly, kinds may be defined satisfactorily without appeal to microstructure, as Lewis argued in the case of acids.[38] Secondly, when there is controversy over the extension of a kind term due to disagreement about what is central, such disagreement is often partly due to shifting assessments of how much promise not-yet-explored

hypotheses and research agendas hold. Since science wouldn't be better off without the latter sort of disagreement, it is hard to see how it would be better off without open texture. And in fact, history of science offers several examples of prolonged conceptual disputes between experts over different characterizations, due to open texture. Such dispute, it seems, hasn't hindered or even necessarily hurt communication, understanding or progress (cf. Cowie 2009, 90-97).

Thus it appears that open texture is simply an ineliminable by-product of how terms are used, both in science and elsewhere. Hence, it shouldn't be surprising that neither causal nor cluster description theories are successful in blocking it, which LaPorte notes – but on the other hand, it is not clear why it should be, as LaPorte says, "unwelcome" (2004, 118). To us it seems a feature rather than a bug. In any case, its pervasiveness fits naturally with a cluster theory for kind terms.

The last few paragraphs were concerned with extensions of kind terms in the actual world. When the issue concerns extensions at all possible worlds, as in the modal arguments of Kripke, the morals still apply. If the earlier use of a term fails to dictate how the term will apply after discoveries revealing inherent vagueness, clearly it doesn't determine application at all possible worlds either. In discussing a different but related topic (modal epistemology), Yablo comments that "grasp of meaning is not a normative crystal ball telling us what modal conclusions are to be drawn from every new empirical finding, however unforeseen or unforeseeable" (Yablo 2000, 120). We concur: applying or refusing to apply a kind term (say, "water") to a hypothetical scenario (Twin Earth) need not betray lack of understanding of the term.

Concluding remarks

We have argued that experimental semantics provides an important source of evidence for semantic theory – in particular when it comes to descriptive semantics, which are clearly empirical theories, but also when it comes to the more theoretically committed metasemantics. Although semantics requires philosophical expertise, the experts cannot ignore the empirical evidence provided by speaker use. Nor can they exclusively rely on their own dispositions to use terms – not, at any rate, if they are interested in the semantics of the shared language.

Although the role of empirical evidence has recently received attention when it comes to proper names, it has received much less attention when it comes to the natural kind terms. Instead, it has generally been assumed that the considerations that apply to proper names apply equally to natural kind terms, and that the intuitive evidence against descriptivism is equally strong. We have questioned this orthodoxy, and argued that the extant empirical evidence does not support a Kripkean semantics for these terms. No doubt, the experimental evidence is still too scant to draw any firm conclusions. However, we have argued, it is safe to conclude that the widely endorsed assumption that natural kind terms cannot be given a descriptivist semantics is nothing but armchair semantics – an empirical theory driven by prior philosophical commitments without support in the available empirical data.[*]

[*] Thanks to Max Deutsch for valuable comments on a draft of this paper.

References

Ben-Yami, H. (1991), 'The semantics of natural kind terms', *Philosophical Studies,* 102, 155-84.

Bird, A. (2009), 'Are natural kinds reducible?', in Hieke, A. and Leitgeb, H. (eds.), *Redcution – Abstraction – Analysis*. Frankfurt: Ontos, pp. 127-36.

Braisby, N., Franks, B., and Hampton, J. (1996), 'Essentialism, word use, and concepts', *Cognition,* 59, 247-74.

Brønsted, J. (1928), 'Acid and Base Catalysis', in Sandved, K. H. and LaMer, V.K. (trans.), *Chemical Reviews* V, 231–338.

Burge, T. (1986), 'Individualism and psychology', *Philosophical Review* 45, 3-45.

Burge, T. (2007), 'Introduction', in Burge, T., *Foundations of Mind*. Oxford: OUP, pp. 1-31.

Cappelen, H. (2012), *Philosophy Without Intuitions.* Oxford: OUP.

Cowie, F. (2009), 'Why isn't Stich an elimiNativist?', in Murphy, D. and Bishop, M., *Stich and His Critics*. Malden, MA: Wiley-Blackwell, pp. 74-100.

Deutsch, M. (2009), 'Experimental philosophy and the theory of reference', *Mind and Cognition*, 24(4), 445-66.

Devitt, M. (2011), 'Experimental semantics', *Philosophy and Phenomenological Research,* LXXXII, 418-35.

Dummett, M. (1981), *Frege: Philosophy of Language*. Cambridge, Massachusetts: Harvard University Press.

Gelman, S.A. and Wellman, H.M. (1991), 'Insides and essences: early understandings of the non-obvious', *Cognition*, 38, 213-44.

Genone, J. and Lombrozo, T. (2012), 'Concepts possession, experimental semantics, and hybrid theories of reference', *Philosophical Psychology*, 25, 717-42.

Griffiths, P. and Neumann-Held, E. (1999), 'The many faces of the gene', *BioScience*, 49 (8), 656-62.

Hacking, I. (2007), 'Putnam's theory of natural kinds and their names is not the same as Kripke's', *Principia* 11, 1-24.

Häggqvist, S. and Wikforss, Å. (2008), 'Externalism and a posteriori semantics', *Erkenntnis* 67, 373-86.

Hampton, J. (1995), 'Testing the prototype theory of concepts', *Journal of Memory and Language,* 34, 686-708.

Hampton, J., Estes and Simmons (2007), 'Metamorphosis: essence, appearance, and behavior in the categorization of natural kinds', *Memory and Cognition,* 35, 1785-1800.

Horvath, J. (2010), 'How (not) to react to experimental philosophy', *Philosophical Psychology* 23, 447-80.

Ichikawa, J. (2009), 'Explaining away intuitions', *Studia Philosophica Estonica* 2.2, 94-116.

Jylkkä J., Railo, H. and Haukioja, J. (2009), 'Psychological essentialism and semantic externalism: evidence for externalism in lay speakers' language use', *Philosophical Psychology*, 22, 37-60.

Kalish, C.W. (1995), 'Essentialism and graded membership in animal and artifact categories', *Memory and Cognition,* 23, 335-53.

Keil, F. (1989), *Concepts, Kinds and Cognitive Development*. Cambridge, Massachusetts: MIT Press.

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

LaPorte, J. (2004), *Natural Kinds and Conceptual Change*. Cambridge: CUP.

Lewis, G. (1938), 'Acids and Bases', *Journal of the Franklin Institute* 226, 293–313.

Luder, W. and Zuffanti, S. (1946), *The Electronic Theory of Acids and Bases*. New York: John Wiley & Sons.

Machery, E. (2009), *Doing Without Concepts*. Oxford: OUP.

Machery, E. (2011), 'Expertise and intuitions about reference', *Theoria*, 73, 37-54.

Machery, E. (2014), 'What is the significance of the demographic variation in semantic intuitions?', in Machery, E. and O'Neill, E. (eds.), *Current Controversies in Experimental Philosophy*. New York, N.Y.: Routledge, pp. 3-16.

Machery, E., Mallon, R., Nichols, S. and Stich, S. (2004), 'Semantics, cross-cultural style', *Cognition,* 92, B1-B12.

Machery, E., Mallon, R., Nichols, S. and Stich, S. (2013), 'If folk intuitions vary, then what?', *Philosophy and Phenomenological Research,* LXXXVI, 618-35.

Machery, E., Olivola, C and De Blanc, M. (2009), 'Linguistic and metalinguistic intuitions in the philosophy of language', *Analysis*, 69, 689-94.

Mallon, R., Machery, E., Nichols, S. and Stich, S. (2009), 'Against arguments from reference', *Philosophy and Phenomenological Research,* LXXIX, 332-56.

Malt, B. (1994), 'Water is not H_2O', *Cognitive Psychology*, 27, 41-70.

Martí, G. (2009), 'Against semantic multiculturalism', *Analysis,* 69, 42-48.

Needham, P. (2010), 'Microessentialism: what is the argument?', *Noûs*, 45, 1-21.

Needham, P. (2012), 'Natural kind thingamajigs', *International Studies in the Philosophy of Science*, 26, 97-101.

Putnam, H. (1962), 'The analytic and the synthetic', in Feigl, H. and Maxwell, G. (eds.), *Minnesota Studies in the Philosophy of Science,* vol. 3. Minneapolis, Minnesota: University of Minnesota Press, pp. 358-97.

Putnam, H. (1975), 'The meaning of "meaning"', in *Philosophical Papers vol. 2: Mind, Language, and Reality.* Cambridge: CUP.

Rips, L. (1989), 'Similarity, typicality and categorisation', Vosniadou and Ortony, A. (eds.), *Similarity and Analogical Reasoning*. Cambridge: CUP, pp. 21-59.

Sarkar, S. (1996), 'Biological information: a skeptical look at some central dogmas of molecular biology', in Sarkar, S. (ed.), *The Philosophy and History of Molecular Biology: New Perspectives*. Dordrecht: Kluwer, pp. 187-231.

Soames, S. (2002), *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity.* Oxford: OUP.

Stalnaker, R. (1997), 'Reference and necessity', in Hale, B. and Wright, C. (eds.), *A Companion to the Philosophy of Language.* Oxford: Blackwell, pp. 534-54.

Stanford, K. and Kitcher, P. (2000), 'Refining the causal theory of reference for natural kind terms', *Philosophical Studies,* 97, 99-129.

Stanley, J. (1997), 'Names and rigid designation, in Hale, B. and Wright, C. (eds.), *A Companion to the Philosophy of Language*. Oxford: Blackwell, pp. 555-85.

Waismann, F. (1945), 'Verifiability', *Proceedings of the Aristotelian Society* (supplementary volume 19), 119-50.

Wikforss, Å. (2008), 'Semantic externalism and psychological externalism', *Philosophy Compass,* 3 (1), 151-81.

Williamson, T. (2011), 'Philosophical expertise and the burden of proof', *Metaphilosophy*, 42, 215-29.

Yablo, S. (2000), 'Textbook Kripkeanism and the open texture of concepts', *Pacific Philosophical Quarterly,* 81, 98-122.

[1] The study involved undergraduate participants from Rutgers and The University of Hong Kong, all fluent in English. Among other probes, they were presented with this vignette and question, modeled on Kripke's example of "Gödel" and "Schmidt" in *Naming and Necessity* (Kripke 1980, pp. 83-92):

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt", whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, he has been

known as the man who proved the incompleteness of arithmetic. Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel. When John uses the name "Gödel", is he talking about:

(A) the person who really discovered the incompleteness of arithmetic? or
(B) the person who got hold of the manuscript and claimed credit for the work? (MMNS 2004, B6)

Answers consonant with (B) were scored as 1; answers consonant with (A) as 0; the mean score for "Western" participants was 1.13 (SD = 0.88) and for Chinese participants 0.63 (SD = 0.84). Similar results were obtained for a Gödel case involving a historical Chinese astronomer (MMNS 2004, B9-B10).

[2] The closest MMNS (2009) gets to outright rejection of the method is this passage: "Our data seem to show that two individuals can belong to two distinct intuition groups despite evidently speaking the same dialect (because they speak the same language, belong to the same culture and have much the same socio-economic status). Faced with this variation, it is very tempting to abandon the assumption that intuitions about reference provide evidence about reference all together [sic]." (MMNS 2009, 345). This passage occurs in the context of an extended attack on "referential pluralism" – the idea that "differences in intuitions about the reference of a word *t* (or a class of words *T*) indicate that *t* (or every member of *T*) refers differently for different groups" (MMNS 2009, 343). What the authors call pluralism thus incorporates the method of cases, but

their attack seems more concerned with pluralism itself than with its method, and ends with a plea to abandon it, rather than the method of cases. We are not sure how to reconcile the quoted passage with the authors' stated reluctance to drop that method. Moreover, Machery has recently explicitly argued that laypeople "are likely to produce true judgments about the reference of proper names in actual and possible cases" and that therefore, "the resulting judgments plausibly constitute evidence" (Machery 2014, 14). This looks like a defence of the method of cases.

[3] For discussion, see Stalnaker (1997, 535-536).

[4] Burge's metasemantics, for example, is externalist and yet he rejects Millian semantics and endorses what seems to be a version of traditional descriptivism (see Burge 1986 and 2007, 1-31). For a discussion of externalism and the relation between semantics and metasemantics see Wikforss 2008.

[5] For example, one might appeal to a principle of "best fit" (such as Davidson's principle of charity) or to the principle that provides the best teleological explanation (as Millikan does) or to assymmetric-dependency considerations (as Fodor does), etc.

[6] See for instance Stanley (1997). Stanley stresses that the discovery that proper names are rigid designators can be accommodated by some versions of descriptivism, such as the "actualized" description theory, and that Kripke does not conclude, from rigidity alone, that Millianism is correct. The philosophical significance of the discovery that names are rigid designators, Stanley argues, should therefore be a matter of controversy. That is, in our terminology, even if the weak semantic theory is not a matter of controversy (such as names being rigid designators) the strong semantic theory is, bringing to bear all sorts of philosophical considerations.

[7] Moreover, the thesis is hostage to empirical assumptions about the kinds in question, (cf. Häggqvist and Wikforss 2008). We return to this below.

[8] See also Machery (2011, 38), where he stresses that theories of reference "belong squarely to metasemantics".

[9] In the initial experiment the participants are invited to respond to questions concerning what the speaker, in the story, uses the name "Gödel' to refer to. The trouble with this way of asking the question, Martí argues, is that it does not test the right kind of intuitions. By making the question explicitly meta-linguistic, asking *what the term refers to*, what is tested is not how the participants use their names, but which theory of reference they think is correct: "MMNS 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" (Martí 2009, 44).

[10] Of course, it is an empirical question to what extent the meta-linguistic judgments of speakers track their first-order use. Thus, in a response to Martí's criticisms, Machery, Olivola & De Blanc (2009) have carried out experiments testing whether metalinguistic intuitions are in agreement with linguistic intuitions. They conclude that these intuitions are in agreement and that the variation between cultures and within cultures remains when the test questions are not meta-linguistic. However, Martí (forthcoming) argues that that new experiments still do not collect the right kind of data since they too prompt the subjects to *reflect* on their use of "Gödel", rather than require them to *use* the name.

[11] For a similar line of reasoning see Cappelen (2012).

[12] The same holds for Evans's famous "Madagascar" counterexample to the causal-historical theory. Since it is common knowledge that "Madagascar" refers to the island,

and not to a portion of the African mainland, Evans did not need to provide evidence for the counterexample, but it is nevertheless an empirical fact that "Madagascar" has this reference. For a related point see Machery's response to Deutsch (2014: 9).

[13] Cf. Ichikawa: "Suppose someone did a survey and discovered that the distribution of people who believed the earth was more than one million years old correlated with certain demographic variables. … I think that the obvious thing to say is that we've discovered that members of a certain demographic group are not reliable judges about the age of the Earth" (2009, 110-111). And Williamson: "[W]e do not expect physicists to suspend their current projects … on the basis of evidence that undergraduates untrained in physics are bad at conducting laboratory experiments" (2011, 217).

[14] This is why semantics involves metaphysics in the Putnam-Kripke theory. Since both hold that the extension of "tiger" across possible worlds is determined not by the psychological states of the speakers, but by the underlying nature of the actual animals picked out, an answer to the metaphysical question (what is the essence of tigers?) is required for an answer to the semantic question (what is the extension of "tiger"?).

[15] This is not to deny that there may be any number of modal propositions that experts as well as non-experts do agree on.

[16] Cf. Machery (2011). Machery appeals to experimental data suggesting that philosophers are more likely to have Kripkean intuitions than linguists working in discourse analysis, historical linguistics and sociolinguistics (2011, 48).

[17] As is often noted, Kripke denies defending any real theory and so he may not in fact endorse the Millian theory. However, he explicitly states that he agrees with Mill on proper names but disagrees with him on general terms, since he thinks that proper names and general terms should be understood along the same lines (Kripke 1980, 127).

[18] See for instance Putnam (1962).

[19] See Ben-Yami (1991, 161), who notes that ”iron pyrites only have a *faint* resemblance to gold”. See also LaPorte (2004, 161).

[20] For a fuller discussion of this point, see Stanford and Kitcher (2000, 105 and *passim*).

[21] It might be suggested, as Max Deutsch did to us, that Kripke’s claim here is not an empirical claim but rather the metaphysical claim that the tiger-looking things are a different species. However, the metaphysical claim, in itself, does not support the semantic claim that Kripke wishes to make, i.e. the claim that the tiger-looking things are not in the extension of our term “tiger”. In order to support the semantic claim, Kripke needs evidence that “tiger” is used in such a way that it tracks underlying features rather than observable features; this is precisely why he makes a claim about what “we would conclude” – an empirical claim about our reactions which, in fact, does not seem supported by the empirical evidence (see section 3.2 below).

[22] And although the notion does make sense in the case of substances, it is utterly unclear what Kripke, Putnam or subsequent semantic theorists mean by the term (Needham 2010).

[23] Dummett was early to reject it, writing that ”Kripke’s efforts to show that that by which we originally identify the species or the substance might not be true of it at all … are bizarre and quite unconvincing” (1981, 146).

[24] Precisely what is meant by “concept” varies, and it is a matter of dispute how the psychological notions relate to the philosophical notion of a concept (cf. Machery 2009).

[25] The type of psychological essentialism that is often tested is so-called “placeholder essentialism”, according to which subjects need not have any specific beliefs about the actual essence of the kind, but merely the belief that the kind has some essence or other,

consisting of deep properties that cause the macro-level properties that are typical of the category.

[26] For further experiments that show similar results see Malt 1994 and Hampton, Estes and Simmons (2007). In the latter experiment the support for essentialism was even weaker, where only 7 out of 110 participants categorized fully in accordance with essentialism.

[27] Indeed, as we shall suggest below, there are reasons to think that the scientists do not use kind terms this way either.

[28] Hacking (2007) notes that there are differences between Kripke's and Putnam's theories, but these are small enough to be finessed here (as they usually are in the literature).

[29] JRH (2009) also use a third-person version of the scenario, asking the participant to answer the same two questions concerning the expert's judgments. Moreover, there was a positive version of each scenario, where an instance first thought of as not belonging to a kind turns out to belong to it, after all. The description here is simplified in several respects in order to focus on the central issues.

[30] Genone and Lombrozo say that "participants were asked whether or not the concepts two characters associate with a common word share the same reference" (2012, 723). This way of glossing their question may be problematic, but there is not room to discuss this here.

[31] The second experiment also tested whether information about the learning history of the characters influenced judgements – it didn't, which, as Genone and Lombrozo note, "suggests that the findings in experiment 1 were not an artefact of having failed to specify a learning history" (2012, 730).

[32] The beliefs at issue were always four, and experiment 1 tested for falsity of one, two, three, and four false beliefs separately; experiment 2 tested for falsity of one and four false beliefs. In both cases, a character's having four false beliefs was thus tantamount to being wrong about everything stipulated to hold of the kind in question.

[33] LaPorte points out that speaker dispositions were supposed to motivate "the conviction that microstructure trumps superficial properties is reference" (2004, 98), and that it would be ad hoc to disregard the evidence provided by the case of "jade". As we have noted in section 2, dispositions constitute important evidence for semantic theories (at least when manifested), but error, and even dispositions to err, must be allowed for.

[34] This is the term LaPorte, following Waismann (1945), uses for hidden and unforeseeable vagueness in a general term.

[35] Or other properties thought to serve a unifying role, such as geneaology in the case of cladistic taxonomy.

[36] As Stanford and Kitcher admit, "some natural kind terms have their references fixed through the use of descriptions that are not attempts to pick out "inner constitution' or 'underlying structure'" (2000, 124).

[37] For "gene", see Sarkar (1996), Griffiths and Neumann-Held (1999).

[38] For an extended argument that even chemical kinds are in fact perfectly well individuated in terms of macroscopic properties, see Needham (2010).