13
Theories of reference
What really is the question?

Jaakko Kuorikoski

In an essay in the honour of Michael Devitt, Panu Raatikainen (2020) takes up the challenge of answering to recent criticisms against the causal-historical or "new" theory of reference by advocates of the description theory. One of the key motivations for reviewing these various critiques, and the original claims by Kripke and others, is to see whether the parties in the debate even share a common understanding of what is the central question that a philosophical theory of reference is supposed to answer. Hence the title of his essay: "Theories of Reference: What is the Question?" In this essay in honour of Raatikainen, I intend to revisit that very question, as I believe more can be said about this very important, but often neglected meta-question. In doing so, I place this article within this particular causal historical chain of reference, and hopefully this will ensure that whatever it is that Devitt and Raatikainen were writing about, this essay will at least succeed in discussing the same topic: what should an adequate philosophical theory of reference be able to accomplish? As both Devitt and Raatikainen are avowed naturalists and take philosophical semantics to be a part of an encompassing empirical account of language, I will approach this question from the perspective of philosophy of science.

After briefly revisiting the history of the philosophical debate on reference from Mill to the emergence of the new theory of reference, Raatikainen answers his titular question in the following way:

Kuorikoski, J. (2025). Theories of reference: What really is the question? In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 152–161).
https://doi.org/10.61201/h5fbdk64

> *Main question: In virtue of what does a referring expression refer to whatever it in fact refers to*? (Raatikainen 2020, 73)

As an answer to a question presented in the very title of an article, this might strike the reader as somewhat underwhelming, as one would think that there is not much disagreement or confusion about whether this really is the main question of philosophical theories of reference. Raatikainen quotes the key players, such as John Searle, Devitt, and William Lycan, who seem to be, more or less, in agreement that yes, this is the common task to be answered by philosophical inquiry into reference. The onus was originally on the reference of proper names, but as the hypothesis that the semantics of natural kind terms behave similarly to that of names gained traction, this seemingly semantic question began to acquire much bigger epistemological and metaphysical stakes.

Raatikainen first goes through the subtle shifts in the history of theory of reference, from discussion of meaning of proper names to reference as such, noting that none of the original theorists intended the theory to be a fully general theory of reference nor meaning, i.e., answer the main question for all possible expression types. He then argues that the modern versions of descriptivism, such as causal descriptivism and metalinguistic descriptivism, are not really up to the task of satisfactorily answering the main question. I will not review or assess Raatikainen's convincing rebuttals against the descriptivist proposals, and the reader is invited to look at the thorough and knowledgeable argumentation from the source. My intention is to take a step back and ask what kind of a question the main question is supposed to be in the first place, and whether the standard philosophical methodology of imaginary counterexamples is really fit for the task of answering it.

Before we start analysing the question in more detail, it is worth pointing out the broader philosophical stance shared by the causal-historicists and the descriptivists: that there is such a thing as the reference relation and that this relation has some important *explanatory* role in the big picture of understanding linguistic communication and perhaps even of our epistemic lot in the world. In contrast, different deflationist accounts of reference deny that there is a substantive relation between a word and its referent to which the concept of reference itself refers to, and that the meaning of 'reference' ought to instead be understood in some purely intra-linguistic way.

Now let us get back to the main question: In virtue of what does a referring expression refer to whatever it in fact refers to? There are at least two points in need of clarification here. What is the nature of the 'in virtue of' relation and what is the nature of the putative 'fact' of referring? What kind of an explanation is the theory of reference supposed to provide and what kind of a phenomenon is it that we are trying to explain?

Disregarding Kripke, at least both Searle and Devitt have stated that their accounts are to be a part of a fundamentally empirical understanding of the phenomenon of language. Especially Devitt has been very explicit about his stern commitment to (meta)philosophical naturalism. He is a self-described card-carrying naturalist and has published a number of important papers attacking the possibility of a priori knowledge (Devitt 2011). More specifically, he has defended at length a thoroughly naturalist methodology for semantics, which includes philosophical (fundamental) semantics as a key element (Devitt 1996). This stance is, I take it, also shared by Raatikainen. An important constraint in clarifying the above questions is therefore that the explanation and the phenomenon ought to be, if not identical with, then at least continuous with the kinds of explanations and phenomena investigated by empirical linguistics, psychology and the like.

"In virtue of"

Let us start with the first item of clarification: what kind of in-virtue-of-relation is at play here? All the key authors, including Raatikainen, insist that the theory of reference ought to be *explanatory* and that the relation thus carries explanatory weight. Again, Devitt is exceptionally clear in formulating the main question in explanatory terms: "The central question about reference is: In virtue of what does a term have its reference? Answering this requires a theory that explains the term's relation to its referent" (Devitt 1998). Furthermore, one of Raatikainen's key arguments against the adequacy of causal and metalinguistic descriptivisms is that they do not offer adequate explanations of reference.

This plea for explanations is not really surprising, as the explanatory commitment is the key feature distinguishing substantive from deflationary accounts of reference. For example, one of the main claims put forward by Brandom (1994), a deflationist about reference, is that representational vocabulary, including the concept of reference, is not itself explanatory, but instead an expressive and explicative metavocabulary. According to Brandom, stating that "'Moo Deng' refers to a baby pygmi hippo" is not to refer to any independently existing relation between the referent and the name explaining its meaning and use, but an act of simultaneously summarizing and instituting a set of inferential commitments and entitlements involving Moo Deng. However, this general claim is a core aspect of the whole Brandomian picture of language and Brandom does not provide any *specific* arguments against the possibility of an explanatory account of reference in particular. Next, I will consider what kind of an explanation the causal-historical theory aims to provide. Together the commitment to an explanatory account of reference and metaphilosophical naturalism mean that the main question ought to be analysable as a scientific explanation, broadly understood. At least to my knowledge no one has seriously asked this question using standard conceptual tools from the philosophy of explanation.

I will start with the assumption that the intended *explanandum* is the fact that a particular expression denotes an object in the world ('Moo Deng' refers to Moo

Deng), and the *explanans* the chain of causally linked utterances of the expression starting from the baptism event.

Even though the very name of the theory refers to causality, let us first quickly discard the possibility that the explanation offered by the causal historical theory could itself be causal in nature. The first objection to this idea is that causal explanations in general are not answers to questions of the type "in virtue of what?". The relata of causal explanations are typically events, whereas here the *explanans* is the whole of the causal historical chain and the *explanandum* the property of an expression (type). Another possibility is that the surface form of the explanation-seeking question is misleading, and that the idea is that the baptism event causally explains the reference at the time of the utterance via the causal chain. An immediate problem with this suggestion is that causal explanation is transitive only in the special case that all the implicit contrast classes in the sequence of explanations line up nicely. Even if we charitably thought that encountering an instance of an expression could in some exceptional circumstances act as a sensible causal explanation of another use of that expression, the idea of a chain of such explanations is implausible.

A more promising suggestion is that the causal historical chain is constitutive of the property of the expression denoting a specific object in the world. This interpretation is also strongly suggested by many philosophers who explicitly state that the causal-historical chain is *the mechanism* in virtue of which the expression has the property of denoting a specific object. For example, Kaplan distinguishes between the way in which an individual is represented from "the mechanism that determines what individual is represented [reference]." (Kaplan 2012, 167, quoted in Raatikainen 2020) and even the Stanford Encyclopedia entry on reference states that the third central question of the theory of reference is "What is the mechanism of reference? In other words, in virtue of what does a word (of the referring sort) attach to a particular object/individual?" (Michaelson 2024). Conceptualizing the causal historical chain as a mechanism also chimes well with the naturalist ambition, as discovering mechanisms is something that the empirical sciences are supposed to be all about.

The problem here is that the putative causal historical chains between baptism events and subsequent uses of an expression do not look or behave anything like other explanatory mechanisms in the sciences. First, the causal historical chain is curiously distributed and extended both in space and especially in time. Let us take a paradigm empirical constitutive explanation of a property or a disposition by its realizing mechanism (in a very broad sense), such as the explanation of the brittleness of an object by its chemical and structural make-up. Here the explanandum is physically and temporally co-extensional with the explanans. The chemical structure in the here and now explains the disposition in virtue of there being a synchronic ontic dependency between the structure and the disposition (e.g. Ylikoski 2013). To be fair, examples of mechanistic explanation in the social science can be more diffuse both in space and in time, as they might involve relational properties, long-term equilibria and the like (Kuorikoski 2009). A particular market mechanism can balance supply and demand of assets or goods with market participants often literally from all over the world and with transactions taking place in dramatically different timescales. Nevertheless, the way in which a property of a (type of) expression here and now would depend constitutively on things that took place hundreds or even thousands of years ago, often in far-away places, is another thing altogether. Typical mechanisms are also relatively stable configurations in which the organization of the parts has an important explanatory role with regard to the property of the whole (Machamer et al. 2000), whereas historical chains of utterances are presumably highly contingent and the pattern of "reference borrowing" does not seem to have any systematic explanatory role.

Perhaps the sensible stance is to take the mechanism-talk as purely metaphorical and admit that the explanations offered by theories of reference are more distinctly philosophical. Clearly a more natural way of understanding the in-virtue-of relation is in terms of *grounding* and theorists of grounding mostly agree that either grounding simply is a form of explanation (e.g., Fine 2012), or alternatively serves as the metaphysical determination relation grounding philosophical explanations (e.g., Schaffer 2016). Although grounding theorists routinely lump many cases of empirical explanations which I would rather call constitutive (e.g. that the bowl's brittleness is grounded on the ionic bonds of its constituent atoms, see Kuorikoski 2012) as cases of grounding, I will restrict my discussion to more conceptual or metaphysical dependencies, as the possibility of constitutively explaining reference in the empirical sense was already dismissed above. The property of referring would thus depend on the causal-historical chain in the same sense as moral and aesthetic facts (if there are any) may depend on non-normative facts about acts and objects of art respectively, truth putatively depends on truth-makers, and essential properties on essences. Such explanations are distinguished, among other things, by implying stronger modality than mere nomological necessity.

The problem for naturalists like Devitt and Raatikainen is that such philosophical grounding explanations are also distinguished by the fact that they have little or nothing to contribute to empirical theories. Whether or not normative properties are grounded in non-normative properties is, at least arguably, inconsequential to any empirical theory of human behaviour, as normative properties do not have any causal power over such matters. Whether or not the redness of a particular colour is grounded in its maroonness is, arguably, pretty much irrelevant for chemistry, optics or neuropsychology of colour perception. Although grounding claims may well have some other, perfectly legitimate, cognitive roles, this interpretation would thus put a serious dent in the naturalist hope that the theory of reference would ultimately serve an explanatory role in a comprehensive empirical theory of language. Furthermore, if the theory of reference were to be a part of such a theory, the *explanandum* itself ought to correspond to an empirically ascertainable phenomenon. Next, I will turn my attention to what kind of fact this might be.

"In fact refers to"

For the advocate of a substantive theory of reference, not only is the causal historical chain supposed to be explanatory of what expressions in fact refer to, this fact about reference itself is also taken to have explanatory value. An obvious instance of this is the idea that reference is part of the meaning of at least some expressions and meaning, whatever it may be, ought to be explanatory of human behaviour (cf. Raatikainen 2020, see also Devitt 1996, ch. 2). If the theory of reference is to be a part of the empirical theory of language, reference ought to be not just a fact, but an empirical fact capable of being investigated by empirical means. Moreover, if these facts were to have explanatory power for human behaviour, they ought to correspond to some robustly *causal* phenomenon (Devitt 2011, 429).

In some sense it seems almost absurd to even question whether matters of fact about reference exist. The point of the whole business of language is presumably to communicate claims about the world, so surely some expressions are really about the world. Only someone who has seriously messed up her worldview with philosophy could deny that there is no such thing as (successful) reference. But it is one thing to admit the reality of linguistic representation and another to claim that there is such a thing as the reference relation. For example, a deflationist like Brandom certainly does not deny that we, as discursive beings, routinely refer to objects with our words – only that the sentence "'Moo Deng' refers to Moo Deng" does not itself refer to a special explanatory relation between the name and its bearer.

It is also clear that referring is, at least in some sense, an empirical phenomenon because it, or at least something closely related to it, actually *has* been empirically studied. Much discussed survey studies in experimental philosophy have claimed to show that there is significant cross-cultural variation in semantic intuitions about reference. Machery et al. (2004) claim that their empirical survey shows that East-Asians have more descriptivist semantic intuitions whereas Westerners think more along the lines of the causal-historical theory. Machery et al. further hypothesize that this difference is linked to a broader cross-cultural cognitive difference between East-Asians and Westerners in that East-Asians' categorization judgments depend more on similarity judgements whereas Westerners focus more on causality. In principle, there should be nothing mysterious about this, as for example grammatical intuitions are known to vary across different linguistic groups.

Max Deutsch (2009) has criticized these studies for falsely portraying the theory of reference (and philosophy of language in general) as relying on "the method of cases" tested against semantic intuitions in the first place. According to Deutsch, the theory of reference "makes predictions" directly about terms and their referents, about *semantic* facts, not about the intuitions of laymen or philosophers. Intuitions about reference and reference are, according to Deutsch, different things. The semantic fact that 'Madagascar' now refers to an island in the Indian Ocean can easily be ascertained by simply opening the Atlas. There is no need to survey any intuitions. Deutsch therefore is clearly committed to the idea of the reference relation as a robust phenomenon existing independently of our intuitions about it. In contrast, Devitt agrees what the theory of reference does resort to semantic intuitions as primary evidence, but argues that the empirical evidence presented by Machery et al. is simply not strong enough. His reasons are that intuitions about hypothetical cases are not as strong evidence as those about humdrum examples, that the intuitions relevant for the case against descriptivism are really metaphysical, not referential in nature, and that philosophers' intuitions really are better evidence than those of the common folk. (Devitt 2011) Although I am not convinced by these counterarguments claiming that the surveyed intuitions are not of the right kind, I will not dwell more on the matter here, as my interest is on the use of semantic intuitions in general.

An important principle of empirical research is that phenomena ought to be multiply measurable by mutually independent means of determination. It is only by triangulating with different independent methods that we can ensure that any putative result is real and not an artifact of any particular method. (Kuorikoski and Marchionni 2016; Wimsatt 2007) The crucial question now is, what other empirical means we really have of deciding whether 'Madagascar' really refers to an island in the Indian Ocean or still to a part of the mainland of Africa, and would do so in a way which would be *independent* of our semantic intuitions about the matter? It is important to clarify here that by semantic intuitions I do not only refer to intuitions in the specific (and perhaps proper) sense of private, immediate, pre-theoretical judgements, but more broadly to also include considered and public interpretations about what people take other people to mean with their words. These interpretations also encompass such things as the Atlas with the depiction of Madagascar in it. Is there any other way of empirically investigating what expressions refer to other than surveying what we take, implicitly as well as explicitly, the said expressions to refer to? At least I have never seen a reference relation and do not know of any empirical methods of detecting or measuring one without first going through our considered judgements about what we think our words refer to. Historians of various ilk certainly produce genealogies of words and concepts, some of which can be philosophically highly enlightening, but such historical narratives are simply further interpretations of historical changes in interpretations of what words mean. One can admit the existence of semantic facts without being committed to the idea that there is a robust empirical phenomenon of the reference relation out there.

Let us finally return to the truism that we use language to communicate claims about the world and that the existence of reference is therefore undeniable. The idea of a substantive reference relation is not solely motivated by the desire to understand the nature of linguistic meaning, but to also understand how linguistic representations are linked to the world outside language. This is also an epistemological worry. An important motivation for believing in substantive reference relations is that these relations could act as semantic hooks anchoring our concepts and beliefs securely into the world. Without such anchors, our system of beliefs is surely doomed to the dreaded frictionless Davidsonian spinning in the void (cf. McDowell 1994, 11).

In philosophy of science, the presupposition that a semantic theory should carry such epistemological weight led to the use of evermore sophisticated theories of reference in the attempt to disarm the pessimistic meta-induction argument against scientific realism. As argued by Stephen Stich and Michael Bishop (1998), this train of thought leads inevitably to some rather bizarre conclusions. If ontological commitments of scientific theories depend on what its terms refer to, and this reference relation is a substantive phenomenon, then we do not really know the ontological commitments of any of our theories until we have found the true theory of reference. But this is plainly mad. Before we elevate the linguistics departments to the highest position in the hierarchy of sciences, we should perhaps rethink the very idea of referential hooks as necessary conditions for our epistemic access to the world. There is plenty of friction with the world even without such contraptions.

Theory of reference as a model of data

If we are serious about the naturalist conviction, what then, remains of the epistemic role of a philosophical theory of reference within an empirical account of language? I definitely do not want to claim that such theorizing is scientifically empty. I suggest that the theory is, in fact, a highly stylized *model of data* in a verbal form. A data model is a representation of data, which highlights some selected systematic features of the data in a cognitively salient manner. In the case of theories of reference, the primary data are the semantic intuitions, understood very liberally as above. The theory thus summarizes a systematic feature in our semantic intuitions: we tend to *judge* or *interpret* people as referring to things in accordance with the tradition of using the expression and with the assumption that at some points in the history of the use of the expression there has been direct interaction with whatever the expression is taken to mean. In fact, this is the very stance that Devitt takes to be the implicit interpretation of the role of semantics by many philosophers, an interpretation he finds deeply mistaken (Devitt 2011). It is important to note, however, that I do not intend to make any sweeping claims about semantics in general, only a suggestion concerning philosophical theories of reference in particular (as, for example, cognitive and computational semantics arguably deal with explanatory relations between language use and cognitive and computational phenomena).

As already Patrick Suppers pointed out in his “Models of Data” (1962), data models are of paramount importance to inquiry. Scientific theories can directly engage with neither phenomena *an sich* nor raw data. Theories explain and are tested by specific (claims about) phenomena, which have to be purposefully and painstakingly distilled from the cacophony of raw data. (Bogen and Woodward 1988) Theories of reference can thus be seen as crystallizations of stylized facts about (a certain aspect) of the phenomenon of language. But what is crucially important to note here is that models of data are not explanatory. They only highlight salient patterns indicative of phenomena, but do not contain epistemic resources to explain those patterns. The explanatory resources lie elsewhere, probably in psychology, socio-linguistics and related fields.

Viewing philosophical theories of reference as data models also partly salvages their standing in the face of the possibility of significant cross-cultural differences in semantic intuitions. Mallon et al. (2009) use the empirical result as grounds to dismiss all philosophical applications of theory of reference: if there is no one correct substantive theory of reference, there can be no arguments from reference. While I agree with the sentiment, it is also important to note that these cross-cultural differences are made much more salient by the vocabulary of theory of reference. Instead of finding ad-hoc arguments to downplay the significance of this data, the true naturalist would welcome the discovery of such interesting systematic differences as potentially important phenomena waiting for an explanation, just not by a philosophical theory of reference.

References

Deutsch, Max (2009): 'Experimental philosophy and the theory of reference', *Mind & Language*, 24(4), 445–466. URL = https://doi.org/10.1111/j.1468-0017.2009.01370.x.

Devitt, Michael (1996): *Coming to Our Senses*, Cambridge: Cambridge University Press.

Devitt, Michael (1998): 'Reference', in E. Craig (ed.), *Routledge Encyclopedia of Philosophy*. London: Routledge.

Devitt, Michael (2011): 'Experimental semantics', *Philosophy and Phenomenological Research*, 82(2), 418–435. URL = https://doi.org/10.1111/j.1933-1592.2010.00413.x.

Fine, Kit (2012): 'Guide to Ground', in Fabrice Correia & Benjamin Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge: Cambridge University Press: 37–80.

Kaplan, David (2012): 'An idea of Donnellan', in J. Almog & P. Leonardi (eds.), *Having in Mind: The Philosophy of Keith Donnellan*, New York: Oxford University Press, 122–175.

Kuorikoski, J. (2009). 'Two concepts of mechanism: Componential causal system and abstract form of interaction', *International Studies in the Philosophy of Science*, 23(2), 143–160.

Kuorikoski, J. (2012). 'Mechanisms, modularity and constitutive explanation', *Erkenntnis*, 77(3), 361–380.

Kuorikoski, Jaakko & Marchionni, Caterina (2016): 'Evidential diversity and the triangulation of phenomena', *Philosophy of Science*, 83(2), 227–247. URL = https://doi.org/10.1086/684960.

Machery Edouard, Mallon Ron, Nichols Shaun & Stich Stephen (2004): 'Cross-Cultural Style', *Cognition*, 92(3): B1-B12. URL = https://doi.org/10.1016/j.cognition.2003.10.003.

McDowell, John (1994): *Mind and world*. Cambridge, MA: Harvard University Press.

Michaelson, Eliot (2024): 'Reference', *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2024/entries/reference/.

Raatikainen, Panu (2020): 'Theories of reference: what was the question?', in Bianchi, A. (ed.) *Language and reality from a naturalistic perspective: Themes from Michael Devitt*, Springer: 69-103.

Schaffer, Jonathan (2016): 'Grounding in the Image of Causation', *Philosophical Studies*, 173(1): 49–100. URL = https://doi.org/10.1007/s11098-014-0438-1.

Suppes, Patrick (1966): 'Models of data', *Studies in logic and the foundations of mathematics* 44: 252–261. URL = https://doi.org/10.1016/S0049-237X(09)70592-0.

Wimsatt, William (2007): *Re-engineering philosophy for limited beings: Piecewise approximations to reality*, Harvard University Press. URL = https://doi.org/10.2307/j.ctv1pncnrh.

Ylikoski, Petri (2013): 'Causal and constitutive explanation compared', *Erkenntnis* 78(2): 277–297. URL = https://www.jstor.org/stable/24010965.