ACTA PHILOSOPHICA FENNICA

Vol. 100

Essays in the Philosophy of Language

Edited by
PANU RAATIKAINEN

HELSINKI 2023

ISBN 978–951–9264–99–8
ISSN 0355-1792

Hansaprint Oy
Helsinki 2023

Table of Contents

Introduction

Philosophy of language has truly become a central subfield of contemporary philosophy. Further, it now has multiple connections with various other areas of philosophy—from the philosophy of science to social and political philosophy and ethics.

Philosophers have always been interested in language. It raises many philosophical questions. Nevertheless, it is customary to begin expositions of the contemporary philosophy of language with Mill and/or Frege. John Stuart Mill (1806–1873) distinguished what he called "connotation" and "denotation"—a distinction somewhat analogous to (though certainly not identical) with Frege's famous sense/reference distinction. Connotation for Mill is a relation between a name (singular or general) and one or more attributes. According to Mill, all general names and most individual names have a connotation, but proper names have none. Consequently, it is now common to label views as "Millian" if they contend that the meaning of a proper name is simply its bearer.

Nonetheless, it was Gottlob Frege (1848–1925) in particular who gave a crucial impetus to the modern philosophy of language. However, the early contemporary philosophy of language—from Frege and Russell to Carnap and beyond—was not primarily interested in ordinary language. On the contrary, these philosophers were part of the ideal language tradition which derived from Leibniz. They had a somewhat dim view of colloquial languages and saw them as problematically ambiguous and vague. Accordingly, they aimed to construe an artificial logical language which would not have such shortcomings. (This tradition is reviewed in the editor's article "Varieties of Ideal Language Philosophy" below. The paper can be viewed partly as a supplement to this introduction.) All the same, these philosophers put forward various

ideas, theses, and arguments which seemed to apply to all languages, whether artificial or colloquial.

One idea that played an important role in Frege's philosophical thought is the so-called context principle: "never ... ask for the meaning of a word in isolation, but only in the context of a propostion." This principle later became central in the thinking of, for example, Wittgenstein, Quine, Davidson, and Dummett. Another of Frege's significant ideas is the principle of compositionality, according to which the meaning of a sentence is exhaustively determined by the meanings of its constituent expressions and the sentence's structure. This principle, in various forms, has also been vastly influential.

Frege's possibly most far-reaching philosophical idea (although somewhat similar distinctions had been made earlier) was his distinction between *the sense* (German "Sinn") and *the referent* (German "Bedeutung") of a referring linguistic expression—roughly the same is referred to in the literature as "intension" and "extension." According to Frege, two names or other expressions can have distinct senses (intensions) even if they refer to the same object (extension). According to Frege, an object can never be accessed directly, without a mediating sense. Every meaningful expression has a sense, although perhaps not always a referent. Frege said unfortunately very little more about what more exactly his senses were supposed to be. However, he stated that a sense is the mode in which an entity is presented—the way the entity is known. It can therefore perhaps be taken, roughly, as a combination of some properties of an object (relevant for the particular way of identifying it). Frege noted in passing that the sense of "Aristotle" might be, for example, the pupil of Plato and teacher of Alexander the Great.

Although Frege's notion of *sense* was arguably more an epistemological than a semantic concept, many of his observations concerning it seemed to apply almost as such to linguistic meaning. It has been popular to consider Frege's arguments (so-called Frege's puzzles) as strong evidence for the view that the meaning of a proper name cannot be—as Mill, for example, had proposed—only the object to which the expression refers.

Along with Frege, Bertrand Russell (1872–1970) played a key role in the formation of the new tradition of linguistic philosophy. Russell later said that he was not really interested in meaning until 1918. All the same, some of his earlier philosophical investigations had a tremendous impact on the development of language-centered analytic philosophy as a whole. The most important was the classical analysis of definite descriptions presented in Russell's classic article "On Denoting" (1905). In Ramsey's words, his analysis formed "a paradigm of philosophy." It seemed that revealing the actual logical structure of a problematic sentence via "logical analysis" dissolved a philosophical puzzle. Perhaps something similar could happen more regularly, some thought.

Russell also suggested that common proper names such as "Socrates" or "Bismarck" *are* actually (definite) descriptions or "abbreviations" of descriptions. Literally taken this is implausible, but it has usually been interpreted to mean that a name is at least synonymous (has the same meaning) with some description—a description that a language user attaches to the name, for example, "Socrates" with "the Greek philosopher who drank hemlock" or "Bismarck" with "the first chancellor of Germany."

\+ + +

The ideas of Frege and Russell influenced the early thought of Ludwig Wittgenstein (1889–1951). However, he went on to develop a view of philosophy which was just as radical as it was original, and it also became highly influential. In his legendary little book *Tractatus Logico-Philosophicus* (1921), Wittgenstein declared that "most of the propositions and questions to be found in philosophical works are not false but nonsensical." What is then left as legitimate philosophy? According to him, philosophy simply is not at the same level as various sciences. In Wittgenstein's view, appropriate philosophy does not involve theories or doctrines at all but is an *activity*—the activity of clarifications of thoughts and especially of the critique of language. It was Wittgenstein's *Tractatus* above all that brought about *the linguistic turn* in philosophy.

The logical positivists of the Vienna Circle were deeply impressed by Wittgenstein's early work, although they interpreted it through colored lenses, with their "scientific philosophy" and radical empiricism. They whole-heartedly welcomed the verdict that traditional metaphysics was not even false; it was actually meaningless. With a certain input from Wittgenstein, the logical positivists contended that unless a declarative sentence is *verifiable* by sense experience, it is *meaningless*. However, from early on, logical positivists such as Carnap distinguished "cognitive meaning" and "expressive meaning." The qualified suggestion now was that unverifiable sentences lack cognitive meaning, but they may still have expressive meaning. Logical positivism became a very influential early trend in the linguistic turn in philosophy.

The austere view on meaningfulness held by Wittgenstein and the logical positivists also had radical consequences for *ethics*. Namely, ethical sentences were judged as strictly speaking meaningless, i.e., lacking any cognitive meaning. This led some philosophers to advocate *emotivism*, which contends that ethical claims are used solely to express emotional attitudes of approval or disapproval and to evoke similar feelings or attitudes in other people (the view is sometimes called "the boo–hurrah theory"). Crude emotivism fell out of favor long ago, but more sophisticated *non-cognitivist* views such as different variants of *expressivism* are seriously considered in ethics, or metaethics, even today (see, e.g., the contribution of Toppinen and Venesmaa in this volume).

Rudolf Carnap (1891–1970) is best known as one of the central figures of the Vienna Circle and logical positivism. Of course, the key ideas of logical positivism have long since been widely rejected. However, Carnap also made various contributions to philosophy, which are at least partly independent of those rejected ideas. For example, his book *Meaning and Necessity* (1947) pioneered the use of the idea of a possible world (which goes back to Leibniz) in analyzing modalities such as necessity and possibility. The idea has proved very fruitful both in formal semantics for modal logics (beginning with Kanger, Hintikka, Kripke, and Montague) and in more informal philosophical considerations, where it has been a helpful heuristic auxiliary tool.

Carnap also developed the idea of *explication* (he borrowed the term from Husserl). That is, he suggested that philosophy should not merely analyze existing meanings and aim to capture the latter with definitions, but also to actively make meanings more precise with new stipulative definitions—to clarify or "refine" meanings. (Carnap's idea of explication is discussed in more detail in the editor's article "Varieties of Ideal Language Philosophy" below.) Carnap's suggestions have directly inspired the idea of *conceptual engineering*. The latter has become quite a popular theme in the present-day philosophy of language but also in social and political philosophy.

Furthermore, in a relatively late article "Empiricism, Semantics and Ontology" (1950), Carnap elaborated his philosophical standpoint with the help of the distinction between questions which are, on the one hand, *internal* and, on the other hand, *external* to a framework. He suggested that metaphysics tends to conflate these questions. There has been again quite a lot of interest in this particular idea of Carnap in recent philosophy (see, e.g., Eklund's article "Carnapian Frameworks Revisited" in the present volume).

\+ + +

The later Wittgenstein, in his *Philosophical Investigations* (1953) and elsewhere, distanced himself from ideal language philosophy and focused on colloquial languages in their variety as they exist "in the wild." His view of legitimate philosophy did not, however, change much. According to Wittgenstein, philosophical problems are to be solved "through an insight into the workings of our language, and that in such a way that these workings are recognized—despite an urge to misunderstand them […] not by coming up with new discoveries, but by assembling what we have long been familiar with." He aphoristically summarized his view thus: "Philosophy is a struggle against the bewitchment of our understanding by the resources of our language." Whether one accepts such a radical view on philosophy or not, Wittgenstein's later philosophy was certainly instrumental in directing the philosopher's attention to ordinary languages.

Among other things, Wittgenstein observed that language is used for a lot more than making assertions and describing the world. In the same spirit, *Speech act theory*, which views linguistic behavior as acts, was then developed especially by John Austin (1911–1960). He distinguished between *constantives*, which are true or false, and *performatives* (e.g., "I promise," "I baptize"), the mere utterance of which amounts to an act that changes the (social) world. Later, Austin rejected this sharp dichotomy and concluded that a single utterance may have both a constantive and a performative aspect; he began to call these aspects *the locutionary content* and *the illocutionary force* of an utterance. He also studied linguistic acts whose aim is to produce some effect (such as conviction, fear, or amusement) in the receiver.

Paul Grice (1913–1988), for his part, distinguished several types of meaning. According to him, a distinction must be first made between *natural meaning* (a reliable sign; e.g., "those spots mean measles"), and *non-natural meaning*. Grice divided the latter in turn into the *conventional meaning* and the *speaker's meaning*. These often coincide, but not always—for example, in the cases of figures of speech and irony. Grice also developed a theory about "implicatures"; that is, he distinguished what a speaker literally says and what the speaker, in a specific context, implicates with her utterance.

The interest in the above-mentioned contributions of Austin and Grice and related ideas has not been restricted to the pure philosophy of language, but they have also had various interconnections with social and political philosophy, for example.

\+ + +

In the 1960s, Saul Kripke (1940–2022), Hilary Putnam (1926–2016), and Keith Donnellan (1931–2015) developed an approach to meaning, referring, and understanding that deviates sharply from the more traditional conceptions. They contended that a normal language user does not necessarily know and associate with an expression any identifying description or "sense" along the lines that Russell and Frege had suggested.

Kripke first approached the issue more technically with the help of the notion of a "*rigid designator*," i.e., an expression which refers to the same entity in all possible worlds (if to anything). He argued that ordinary proper names are such, whereas common descriptions clearly are not—and that therefore names cannot be synonymous with descriptions. Other modal and epistemological considerations supported the same conclusion.

Kripke, Putnam, and Donnellan, partly independently of each other, also developed philosophically powerful *arguments from ignorance and error* against the views that lean on descriptions in the minds of language users. Kripke proposed that a name used by a language user successfully refers to its correct referent instead through a kind of *causal-historical chain*—even if the language user does not know any conditions sufficient to identify that referent. One may even have numerous untrue beliefs about the bearer of the name. This picture has been called by different names: "the causal theory of reference," "the historical chain picture," and "the new theory of reference." The basic insight is that the more traditional descriptivist approaches do not allow for error and ignorance, which are so common to us humans.

Putnam argued (with the help of his famous Twin Earth thought experiment) that two subjects could in principle have internally exactly the same mental states, but the syntactically identical words they use (or think) could nevertheless refer to different items due to the different environment or the context of use. In particular, the exact meaning of a word may be partly unknown to the users of the word; it may mean or refer in virtue of relations that are partly external to the mind. Putnam summarizes his conclusion with his well-known slogan: "meanings just ain't in the head." Accordingly, the view of Putnam and congenial spirits is often called "*semantic externalism*."

In his book *Wittgenstein on Rules and Private Language* (1982), Kripke presented a startling skeptical paradox about meaning which was inspired by Wittgenstein's later reflections on *rule-following*. (Kripke did not claim that this was necessarily the position of the historical Wittgenstein; he explicitly stated that the problem just occurred to him while

reading Wittgenstein.) It has attracted much attention and given rise to a vivacious philosophical discussion.

The problem is, in broad strokes, the following. Suppose a person has applied a word in the past in a certain way—e.g., has applied it to a certain object or certain types of objects. However, there are countless alternative ways she could apply it in the future. Why exactly would one of these be correct and all the others wrong? *The skeptical challenge* now contends that the past cases of application do not unequivocally determine any one way to be correct in the future; it is only an illusion that the person would have followed some definite rule. As a result, it is also an illusion that the word had a determinate meaning in the past. Kripke noted that this conclusion is, nevertheless, completely implausible and even self-refuting.

Kripke examined different possible responses to the skeptical challenge, e.g., appealing to application dispositions or to an entire language community. However, he argued that these only move the problem one step further but do not eliminate it. It is unclear and controversial what Kripke himself ultimately thought about the matter and what he meant by all this—and what we should think about it; the literature on the theme is vast.

At any rate, it has been quite popular to take Kripke's reflections as supporting the conclusion that meanings essentially involve some kind of *normative* element. More recently, however, this assumption of the normativity of meaning has received increasing criticism from philosophers such as Kathrin Glüer, Anandi Hattiangadi, and Åsa Wikforss. These critical arguments have sparked a lively debate and several responses have been in turn presented. The debate is very much ongoing.

The earlier philosophy of language focused predominantly on declarative sentences and relatively neutral, non-evaluative expressions such as proper names, indexicals, natural kind terms, etc. More recently, however, there has been an increasing interest also in normative sentences and evaluative and disparaging expressions. On the one hand, the existing tools of the philosophy of language may help to analyze and also understand the latter. On the other hand, their analysis may create new ideas and advance the philosophy of

language. This is an active topic in the philosophy of language at the moment.

So-called *experimental philosophy* is a 21st-century newcomer in philosophy (although it has certainly had some precursors). Traditionally, philosophy has been typically done, so to say, "in the armchair," through conceptual analysis, thought experiments, and such. Experimental philosophy, in contrast, casts doubt on the reliability of such methods, and conducts instead empirical surveys focusing on laypeople's intuitions and reactions. A famous early (2004) study by Machery and his collaborators strived to show specifically that the by then highly popular new theory of reference is in fact in a sorry state. They suggested that people from different cultures do not share Kripke's intuitions. Devitt, Martí, and Haukioja, for example, have in turn criticized at least the strongest claims of these experimentalists. These are too still very much ongoing debates. (Martí's article in this volume is a brand-new contribution to this debate. Also, Devitt's contribution touches upon this theme.)

\+ + +

The present collection of philosophical articles brings together some of the leading experts in the philosophy of language and different generations of philosophers from around the world. It provides a multifaceted view of some recent work on various aspects of the philosophy of language. Let us briefly review the articles of this volume.

Artificial formal languages played a pivotal role in early analytic philosophy and the philosophy of language. The branch of analytic philosophy which has focused on new formal logic is often called "Ideal Language Philosophy." The first article of the present volume, "Varieties of Ideal Language Philosophy" by the editor, reviews this tradition. Its aim is to shed light on how and why more exactly those influential philosophers gave such an enormously central place to formal languages in their whole philosophical thought. The different ways these thinkers viewed the role of formal languages and their relation to colloquial languages are tracked. As was mentioned above, this paper functions partly

as a further introduction and provides partial historical background for many of the papers that follow.

The next paper is also historical in nature. It compares Descartes and certain recent ideas in the philosophy of language. Although Descartes' theory of ideas is a debated topic among scholars, its relation to Descartes' account of language has not received wide attention. In his paper "Descartes on Language: How Signification Leads to Direct Reference," Jani Sinokki examines the possibility of reconstructing Cartesian semantical theory on the basis of (i) the few remarks Descartes makes about language; (ii) what we know about the kinds of theories of signification in general to which Descartes is committed; and (iii) his interpretation of Descartes' theory of ideas. He suggests that in the light of considerations (i)–(iii), Descartes seems to be committed to viewing thoughts about particulars as singular (Russellian) propositions and to a causal theory about mental contents not unlike Kripke's causal theory of reference. Sinokki argues that this, in combination with a theory of signification, makes for an interesting view about "semantic content" that exhibits many features usually associated with theories of direct reference rather than views often pejoratively called "Cartesian."

In the last few decades, there has been much renewed interest in Carnap's famous distinction between questions internal and external to a framework. In particular, Matti Eklund has, in a series of papers, scrutinized and elaborated on the distinction, paying special attention to what frameworks are to begin with. His contribution "Carnapian Frameworks Revisited" continues this line of work. Gabriel Broughton has criticized Eklund's discussions of Carnap on ontology and put forward his own interpretation of what Carnap's external/internal distinction amounts to. Eklund first argues that Broughton's main claims about Eklund are based on a misinterpretation. He then turns to some issues that are of broader interest. Eklund argues that Broughton's own, potentially interesting interpretation of Carnap's external/internal distinction does not work. And in light of Broughton's discussion, he presents a sharpened version of what he has said about this distinction.

In "Is Semantics Possible?," Putnam connected two themes: the very possibility of semantics (as opposed to for-

mal model theory) for natural languages and the proper semantic treatment of common nouns. Putnam observed that abstract semantic accounts are modeled on formal languages model theory: the substantial contribution is rules for logical connectives (given outside the models), whereas the lexicon (individual constants and predicates) is treated merely schematically by the models. This schematic treatment may be all that is needed for an account of validity in a formal setting, but it does not help to understand how proper and common nouns function in reality (not in models). Putnam then initiated the empirical study of such nouns to indicate, (i), that the popular Frege–Carnap account of them as ("disguised" compound) predicates is empirically incorrect, and, (ii), that they offer a new paradigm for a naturalistic semantics of natural languages. In their article "The Semantics of Common Nouns and the Nature of Semantics," Joseph Almog and Andrea Bianchi take Putnam's program a couple of steps further. First, they investigate the semantics of common nouns and argue that they refer (to kinds), rather than apply by satisfaction/truth to a designation/denotation. Second, Almog and Bianchi point to general results about semantics as a theory whose fulcrum is the reference relation rather than satisfaction in models and validity across them.

In his paper "The Fallacies of the New Theory of Reference: Some Afterthoughts," Gabriel Sandu revisits some of the arguments in "The Fallacies of the New Theory of Reference" by Hintikka and Sandu (1995). The main claim of that paper was that, contrary to what, e.g., Kripke claims in *Naming and Necessity*, there are no expressions in natural language which function as rigid designators (i.e., refer to one and the same individual in all worlds in which the individual exists) and that, if one wants names to function rigidly in some alethic or epistemic contexts, this can be ensured with the help of quantifiers (de re propositional attitudes). In the paper in the present volume, Sandu argues that these claims were not entirely correct.

There have been as of late several experimental studies on the use of kind terms, often with widely different results. Some of those studies report substantial disagreement among participants and even a good number of contradictory responses by individual participants. In her contribution "Ex-

perimental Results on Kind Terms: A Critical Reflection," Genoveva Martí discusses critically these studies and reflects on their impact on the theory of reference for kind terms.

Michael Devitt's article "Type Specimens and Reference" operates in the intersection between the philosophy of biology and the philosophy of language and metaphysics. Alex Levine has alleged the following paradox: "qua organism, the type specimen belongs to its respective species contingently, while qua type specimen, it belongs necessarily." One major concern of Devitt's paper is to argue that the latter necessity, "Levine's Thesis," is false. This argument is based straightforwardly on the words of biologists themselves. There have been previous responses to Levine's paper by LaPorte, Haber, Witteveen, and Brzozowski, which have found the matter much more complicated. Devitt's other major concern is to show that these responses have gone awry because of mistakes about language. Devitt argues, contrary to these philosophers, that we should not use a theory of reference to assess Levine's Thesis. In any case, the causal theory of reference, once developed to include "multiple grounding," does not imply Levine's Thesis. Finally, we should not make any inferences about species identity, and hence about Levine's Thesis, from decisions about nomenclature. He concludes that the engaging debate about Levine's Thesis has been misguided.

It has been suggested by a number of theorists that semantic externalism is at odds with conceptual engineering: the latter is the project of changing the intensions of our words, while the former claims that the intensions of our words are typically dependent on external matters of fact, over which we have limited or no control. In his contribution "Conceptual Engineering for Externalists," Jussi Haukioja discusses the problem and criticizes Steffen Koch's recent attempt at resolving it. He then goes on to present a somewhat similar solution, which is nevertheless crucially different in some of its details. Haukioja suggests that conceptual engineering is far easier to make sense of within a semantic externalist framework than most theorists, including Koch, have assumed.

Several philosophers including Kripke have contended that fictional entities do exist as abstract objects, and fictional names refer to such abstract entities. Kripke and Thomasson compare fictional entities to existing social entities. Kripke

also reflects on fictions inside fictions to support his view. Many philosophers appeal to the apparent fact that we quantify over fictional entities. Such arguments in favor of the existence of fictional entities are critically scrutinized in the editor's article "Fictional Names Revisited." It is argued that they are much less compelling than their proponents suggest and involve various obscurities.

Philosophical theories of various sorts rely on there being robust boundaries between kinds of content. One way of drawing such boundaries is to place them between subject matters, like physics and aesthetics, and the domains of sentences falling within them. Yet, contemporary literature exploring the nature of discourse domains is relatively sparse. The goal of Teemu Tauriainen's paper "Indeterminism about Discourse Domains" is to articulate the core features of discourse domains for them to provide the sought-after explanatory utility of establishing robust boundaries between discursive contents. Analyzing the role that discourse domains have under alethic theories yields valuable information about the ways in which domains subject themselves to being defined and how alethic theories can explain the variability of truth-aptness or truth across sentences from distinct domains. Tauriainen's concluding argument is that because of certain issues with defining domains as unambiguous classes of sentences when individuated on the grounds of topical subject matters, philosophers should consider a commitment to indeterminism about the extensions of fundamental domains. According to this view, although domains can be defined as relatively well-individuated classes of sentences based on topical distinctions, the temporal development of our conceptual frameworks and the phenomenon of mixed content compromise our ability to definitively account for the domain membership of all truth-apt sentences. Such an indeterminacy argument is relevant for all who rely on there being robust boundaries between topically individuated discursive contents.

According to a widely held view, judgments of validity are normative judgments. For instance, conventionalists hold that a logic is a system of rules which, together with a broader framework of linguistic rules, determine which inferences are valid. Since rules have normative or imperative content, to

adopt a system of rules is to make a normative judgment. Similarly, Field's evaluativism about logic involves the view that to adopt a logic is to adopt a policy for reasoning, and since policies have normative content, his view too implies that judgments of validity are normative judgments. In her paper "Is Logic Normative?," Anandi Hattiangadi takes issue with the view that judgments of validity are normative judgments. By appeal to MacFarlane's helpful categorization of logico-normative bridge principles, she argues that there is no plausible account of what the normative content of judgments of validity consist in.

Timothy Williamson famously argued that assertion is constituted either by the knowledge rule or some similar epistemic rule. If true, the proposal has important implications for criticism of assertions. If assertions are analogical to other rule-constituted kinds like games, we can criticize assertions either on external or internal grounds, depending on whether the criticism draws from the necessary norms of assertion or some contingent ones. More recently, authors like Goldberg and MacFarlane have argued against other theories of assertion on the grounds that they cannot explain the possibility of internal criticism for assertions. The article "Constitutive Rules and Internal Criticism of Assertion" by Jaakko Reinikainen raises methodological problems with these arguments. He argues that alternative, non-normative accounts of assertion can also explain the apparent differences in grounds of criticism without assuming that assertion is necessarily governed by some epistemic norm.

The idea that meaningful expressions have conditions of correct application has often been taken to provide intuitive support for the claim that meaning is essentially normative. Failing to explain this normative component in meaning has been seen as a threat to the plausibility of a variety of traditional theories of meaning. The opponents of the normativity thesis accept the existence of semantic correctness conditions but reject normativity arguing that correctness does not imply categorical semantic prescriptions which tell what speakers ought to do independently of their desire to speak the truth or communicate. In his paper "What Could and What Should Be Said? On Semantic Correctness and Semantic Prescriptions," Aleksi Honkasalo argues that semantic correctness can

be understood nonprescriptively. Additionally, if semantic correctness does imply prescriptions, these prescriptions forbid speakers from using expressions appropriate for their communicative intentions and therefore cannot be semantic in nature.

Sarah-Jane Leslie has proposed an influential psychological view of generics. According to her view, generics are products of a primitive psychological mechanism of generalization. Leslie's central claim is that generics do not have compositional truth conditions. They have much looser worldly truthmakers instead. In his article "Rearticulated Psychological View of Generics and Worldly Truthmakers," Pasi Valtonen offers a new perspective on the relationship between generics and their worldly truthmakers. He argues that the rearticulated view accommodates our intuitions about generics better. Rachel Katharine Sterken has argued that (i) Leslie's worldly truthmakers are open to numerous counterexamples; (ii) contrary to Leslie's thought, generics are context-sensitive; and (iii) generics do not express cognitively primitive generalizations. Valtonen proposes that the rearticulated view he develops enables in addition a response to Sterken.

Joonas Pennanen contends, in his article "The Dual Character of Essentially Contested Concepts," that so-called essentially contested concepts should be understood as having a dual character. He aims to show that there are close affinities not only between the ways of employing essentially contested concepts and dual character concepts but also with respect to a third type, i.e., natural kind concepts. In all three cases, concept-users represent a concept in terms of a deeper unobservable property or essence, which enables categorizing an object according to criteria drawn from its concrete features or with reference to its hidden essence. According to Pennanen, the identification of these characteristics is important because (i) they help in answering some open questions about essentially contested concepts; (ii) the shared organization that consists of distinct sets of criteria of application suggests that dual character concepts and natural kind concepts could also be vehemently contested in suitable circumstances; and (iii) a structural resemblance between the three concept types makes the category of essentially contested concepts less for-

eign and thus gives indirect support for a thesis of essential contestability.

In their contribution "Unified Metasemantics for Expressivists," Teemu Toppinen and Vilma Venesmaa discuss a challenge for expressivism in metaethics. According to expressivism, the meaning of normative sentences is explained by their playing a practical role, or by facts about what "desire-like" states of mind normative sentences express. The challenge, which may be called "the problem of diverse uses," is based on the simple observation that while terms such as "good" or "ought" plausibly have a unified meaning across a wide variety of different uses, not all uses of sentences that contain these terms seem to play a suitably practical role. Toppinen and Venesmaa suggest that expressivists can deal with this challenge if they accept two claims. First, expressivism must be understood as a view in metasemantics rather than in semantics. This makes space for the idea that both the practical and the descriptive uses of normative terms might carry the same meaning. The distinctively expressivist claim is, then, that the metasemantic explanation for why the sentences in question have the meaning that they do have is different in the practical and descriptive cases. Second, in order to avoid implausibly disunified metasemantics, the expressivist account must be of a suitable kind. Toppinen and Venesmaa propose that a view called "relational expressivism" allows us to offer a unified enough metasemantics for normative language. According to relational expressivism, normative sentences express relational states of having one's representational beliefs and desire-like states being related in certain ways. On this kind of view, sentences deploying terms such as "good" or "ought" may be taken to express states involving similarly structured representational beliefs across both practical and descriptive uses. This makes it unsurprising that metasemantic explanations that differ in some respects may nevertheless account for the unity of meaning across the different uses of terms such as "good" or "ought."

Panu Raatikainen

Varieties of Ideal Language Philosophy

PANU RAATIKAINEN

1. Introduction

In the honorable tradition of analytic philosophy,[1] it has been common to distinguish two subordinate traditions, "Ideal Language Philosophy" and "Ordinary Language Philosophy."[2] The latter obviously denotes philosophy which focuses on natural languages. The former refers to the kind of philosophy that utilizes artificial formal languages and emphasizes their importance for philosophy. This tradition has often had a somewhat critical attitude toward colloquial languages. Paradigmatic representatives of this approach include Alfred Tarski and especially Rudolf Carnap. What is distinctive in Ideal Language Philosophy is the central role of the language of new formal logic, due to Frege and Russell. The new logic was initially developed to serve as a tool in the philosophy of mathematics, namely, to enable rigorous gap-free inferences and precise definitions in their attempts to reduce arithmetic conclusively to logic. However, it subsequently achieved a much more general and philosophically pivotal role in the tradition at stake here.

In what follows, I aspire to track how new formal logic became so enormously central to early analytic philosophy. I will look at the beginnings of Ideal Language Philosophy in Frege's and Russell's work, very briefly discuss the role of early Wittgenstein, and review the relationship of its key representatives—Carnap and Tarski—to ordinary language, and

[1] The nature and scope of this tradition is a subject of some debate; see Raatikainen 2013a for my own, somewhat unorthodox view.

[2] The distinction, with these very labels, was influentially propagated by Richard Rorty in his widely read introduction to the collection *Linguistic Turn* he edited (Rorty 1967).

by doing so I aim to shed light on why these philosophers gave such an important place to artificial formal languages in their whole philosophy.

There is already a rich scholarly literature on each of these philosophers. I do not want to pretend that anything I am saying here is big news for scholars. However, I think it is fruitful to provide a comparative overview, a synoptic picture, or a lengthwise cross-section, and collate these highly influential and original thinkers and consider how they viewed the role of artificial formal languages in philosophy. We can then see more clearly both differences and some continuity and similarities, as well as how certain ideas about the relations of colloquial languages and artificial formal languages evolved within this tradition.[3] (I must necessarily set aside many interesting details, including various changes of mind made by the philosophers discussed, in order to keep the size of the paper reasonable.)

2. Background: Leibniz, Characteristica Universalis and Calculus Ratiocinator

An important background figure for the tradition under consideration here is the 17th-century polymath, mathematician, and philosopher Gottfried Wilhelm Leibniz (1646–1716). Aristotle and his followers over centuries had assumed that natural languages reflect quite well forms of logical reasoning and other logical relationships, and perhaps even the structure of reality. Leibniz, on the other hand, thought that everyday words do not adequately reflect reality, and that a new artificial language, modeled after algebra and arithmetic, which would undistortedly mirror the reality and its structure, should therefore be constructed.

---

[3] The only overview with a somewhat similar concentration on the ideal language tradition I am aware of is Hylton 2018; however, his emphasis is quite different (the specific Russellian idea of a logically perfect language), and his focus is more on later Quine and even Lewis than on the earlier figures discussed in this paper. Thus, I think that the present paper and Hylton 2018 nicely complement each other. My understanding of Russell specifically here is, though, indebted to an earlier paper by Hylton (2007).

Leibniz put forward the idea of a universal logical ideal language, "a Characteristica Universalis," which would reflect the structure of the whole world without distortion, and the "Calculus Ratiocinator," a precise and comprehensive system of logical reasoning that would facilitate reasoning by making it entirely mechanical and thus enable the derivation of all truths from simple thoughts.[4] The universal language should include, for any simple thought, a sign designating it unambiguously. It would represent all the logical structure of the world. On the one hand, according to Leibniz himself, it is not possible for man to know the latter, so even in Leibniz's view, a perfect universal language would ultimately be impossible for man. On the other hand, Leibniz did have high hopes for the universal language: for example, he believed that his universal language would help to resolve disputes that had been entrenched in the wars of religion between Catholics and Lutherans, among others. Leibniz's more concrete attempts in logic were extensive but far from completed.[5] His ideas, however, were not completely forgotten either, at least in the German-speaking world,[6] where they were kept alive by some of the less well-known thinkers.

3. Frege and his concept-script

Leibniz's vision had begun to come true—insofar as it was at all possible—in the work of German mathematician and philosopher Gottlob Frege (1848–1925), the founder of modern logic, who is generally considered to be the greatest logician since Aristotle.[7] Frege undertook to construct a new logical language that would implement both of Leibniz's ideas: Frege's logical ideal language was intended to be both a uni-

[4] For Leibniz, these were not two separate projects, but two aspects of the broader project of general science; see, e.g., Peckhaus 2004.

[5] For a rather comprehensive overview of Leibniz's work on logic, see Lenzen 2004.

[6] Recall that Leibniz was—although he wrote in Latin and French—German.

[7] A more complete story should certainly also discuss at least George Boole and the Boolean tradition in logic.

versal medium of expression, a "*Lingua Characterica*,"[8] and a system of rules of logical reasoning, a *Calculus Ratiocinator*, as he interpreted them.

Frege referred to the rather little-known German logician Adolf Trendelenburg, who had written earlier a review of Leibniz's idea of universal language (Trendelenburg 1857). In Trendelenburg's text, the ideal language of Leibniz was called "a concept-script," which Frege adopted as the name of his own ideal language. Trendelenburg peculiarly interpreted Kant as a developer of the Leibnizian ideal language project. As is well-known, Kant distinguished sharply conceptual and empirical components of thought. According to Trendelenburg, Leibniz's original goal is impossible to achieve, but he interpreted Kant's distinction as resulting in the more realistic goal of an ideal language: the task is no more to try to represent in an ideal language all the properties of objects, but only the conceptual properties.[9] Frege adopted this picture of the relationship between Leibniz and Kant. He left empirical objects outside his ideal language and focused on the study of formal concepts. Indeed, Frege sometimes used the label "formula language of pure thought" with a definite Kantian ring of his logical ideal language. (See Sluga 1980, Haaparanta 1985, Beaney 1996, Gabriel 2013.)

For Leibniz, thought and perception were distinguished only by the degree of clarity and distinctness. Kant, on the other hand, made a sharp distinction between the faculties of sensibility and understanding. Frege followed Kant by sharply distinguishing between reason as the source of logical knowledge, perception as the basis of empirical knowledge, and *a priori* intuition as the basis of synthetic *a priori* knowledge. For Leibniz the rationalist, after all, all knowledge was in the end *a priori* and analytic, and an ideal language would make it at least in principle possible to achieve all truths. For Frege, in contrast, the use of the envi-

[8] Frege and Trendelenburg (see below) called the Leibnizian idea "Lingua Characterica," not "Characteristica Universalis," as Leibniz himself had named it. The former likely derives from Erdmann's influential edition of Leibniz's works (1839–40) which also employed that formulation.

[9] Trendelenburg in turn cited Ludwig Benedict Trede (1811), who had earlier put forward somewhat similar ideas. Frege certainly knew about Trede at least through Trendelenburg's summaries.

sioned ideal language was much more limited, as it was restricted to form and hence to logic. Frege, on the other hand, followed Leibniz in that he, too, took a quite critical stance toward natural language: he considered it ambiguous, unclear, and contaminated with erroneous (including psychologistic) philosophy, and did not trust it as a basis for logical knowledge. Frege's concept-script was intended as a new universal language logically superior to natural language. The language of his new logic was published in his first book *Begriffsschrift* ("Concept-script"; Frege 1879).

Frege was indeed dissatisfied with the philosophical theories of his time about mathematical truths and our knowledge of them. Frege took as his vocation to reduce arithmetic to logic. In this way he wanted to demonstrate for good, on the one hand, that the various then-fashionable empiricist and psychologistic theories of mathematics were totally wrong and that knowledge in arithmetic is *a priori*, and, on the other hand—contrary to Kant's claim—that arithmetic was not synthetic but analytic. He soon found traditional Aristotelian logic hopelessly inadequate for this program and developed single-handedly modern propositional logic and quantification theory.[10] Frege's view in the philosophy of mathematics that at least arithmetic is reducible to the truths of logic is commonly called "Logicism."[11] This idea, too, was inherited from Leibniz.

Frege's goal was to show that all the truths of arithmetic can be proved on the ground of "laws of thought that transcend all particulars." Frege states in the Preface to his *Concept-script* that in order to prevent anything intuitive from sneaking in imperceptibly, he sought to keep the chain of inferences free of gaps:

> In striving to fulfill this requirement in the strictest way, I found an obstacle in the inadequacy of language: however cumbersome the expressions that arose, the more complicated the rela-

[10] Beaney 2016 contains an accessible summary of the benefits of Frege's new logic.

[11] On the other hand, Frege agreed with Kant on his conviction that geometry is synthetic *a priori*, and as such, is not reducible to logic (which is analytic). Russell (see below), in contrast, advocated all-encompassing logicism with respect to all of mathematics.

> tions became, the less the precision was attained that my purpose demanded. Out of this need came the idea of the present [concept-script]. It is thus intended to serve primarily to test in the most reliable way the validity of a chain of inference and to reveal every presupposition that tends to slip in unnoticed, so that its origin can be investigated. (Frege 1879, 48–49)

Frege's logical notation was intended to express all the content of any judgment that is relevant to the logical reasoning in which it occurs. It is intended to be a tool for assessing the validity of any inference on any subject matter and for preventing any presuppositions from creeping in. Once our inference is expressed in the concept-script, the expectation is that it is a purely mechanical task to determine whether a given inference is valid and free of gaps, or whether it requires a hidden premise. It must be possible to see by examination whether or not a given claim is a logical law and whether the transition from one claim to another follows the logical rules put forward by Frege.

It follows from the above that not everything that can be expressed in natural language can be expressed in Frege's ideal language. Frege says he has chosen to refrain from expressing anything that is irrelevant to the chain of inferences. He calls what his ideal language expresses "conceptual content." Two judgments from which exactly the same consequences can be deduced are said to have the same conceptual content. The intuitive difference between what the words "and" and "but" express in natural language is a classic example of something that his notation cannot express, and Frege himself mentions it in *Concept-script*.

It might be tempting to assume that Frege's concept-script is only a version of natural language from which additional content that would obscure logical connections has been removed. However, this would be a mistake, as it ignores important differences between the purposes of concept-script and natural language. In the preface to his *Concept-script*, Frege writes:

> I believe I can make the relationship of my [concept-script] to ordinary language clearest if I compare it to that of the microscope to the eye. The latter, due to the range of its applicability, due to the flexibility with which it is able to adapt to the most

> diverse circumstances, has a great superiority over the microscope. Considered as an optical instrument, it admittedly reveals many imperfections, which usually remain unnoticed only because of its intimate connection with mental life. But as soon as scientific purposes place great demands on sharpness of resolution, the eye turns out to be inadequate. The microscope, on the other hand, is perfectly suited for just such purposes, but precisely because of this is useless for all others. (1879, 49)

The microscope does not filter out irrelevant details from the images we see. Rather, the sharpness of resolution makes it possible to see what cannot be seen with the naked eye. Frege therefore believes that the concept-script has expressive power that natural language does not have. In other respects, its expressive power is weaker. Like a microscope, an ideal language is perfectly suited to certain needs, but that is why it is also "useless for all others." The concept-script is a device developed for certain specific scientific purposes and should not be condemned, according to Frege, because it is not suited to some other purposes (ibid.).

For scientific purposes, natural language is deficient. However, these logical faults are, according to Frege, necessary for natural language to serve its own purposes. Elsewhere, Frege also compared natural language to a hand:

> The shortcomings [of ordinary language] stressed are rooted in a certain softness and instability of [ordinary] language, which nevertheless is necessary for its versatility and potential for development. In this respect, [ordinary] language can be compared to the hand, which despite its adaptability to the most diverse tasks is still inadequate. We build for ourselves artificial hands, tools for particular purposes, which work with more accuracy than the hand can provide. And how is this accuracy possible? Through the very stiffness and inflexibility of the parts the lack of which makes the hand so dexterous. Word-language is inadequate in a similar way. We need a system of symbols from which every ambiguity is banned, which has a strict logical form from which the content cannot be escape. (Frege 1882, 86)

If we are interested in something that serves the purposes of natural language, then the logical notation of the concept-script is inadequate. It would therefore be a mistake to de-

scribe Frege's logical language as a properly functioning version of natural language. Frege's notation is intended not to be a perfect language but a *logically* perfect language.[12]

As a philosopher, Frege still belonged to the earlier broadly Kantian epistemological tradition to a significant extent, and he did not yet—like Wittgenstein or Carnap later (see below)—put forward any radical general theses about the aims and the scope of philosophy. Still, with hindsight it is difficult not to see the following words in the preface to *Concept-script* as anticipating and grounding what was to come:

> If it is a task of philosophy to break the power of words over the human mind, by uncovering illusions that through the use of language often almost unavoidably arise concerning the relations of concepts, by freeing thought from the taint of ordinary linguistic means of expression, then my [concept-script], further developed for these purposes, can become a useful tool for philosophers. (Frege 1879, 50–51)

4. Russell and logically perfect language

Along with Frege, another early key figure in Ideal Language Philosophy is the influential British philosopher Bertrand Russell (1872–1970). He knew Leibniz's thought firsthand—after all, he had published the book *A Critical Exposition of the Philosophy of Leibniz* (1900). Russell too found Leibniz's original universal language project in its entirety unrealistically ambitious, but he believed in its feasibility in the area of mathematics (Russell 1903a). From Leibniz, Russell also adopted the idea that there is no specifically mathematical method but that mathematics reduces to logical truths, i.e., the core idea of logicism.

Russell reported in retrospect that "the most important year in his intellectual life" was the year 1900, when he attended an international mathematics conference in Paris and in particular heard Giuseppe Peano, an Italian mathematician and logician (Russell 1944, 12). Russell was greatly impressed by the artificial symbolic language developed by Peano,

[12] Some further ideas of Frege which he shared with Russell will be discussed in the next section.

which seemed to him to provide a new powerful method for the study of the foundations of mathematics. Peano also explicitly saw his symbolic language as an extension of Leibniz's program. In the early years of the 20th century, Russell then also delved into Frege's work and it certainly influenced him. However, Russell had already ended up with many of the key ideas independently.[13]

Russell later said that he was not really interested in meaning until 1918. All the same, certain previous philosophical investigations by him had a tremendous impact on the development of analytic, language-centered philosophy as a whole. The most important was the classical analysis of definite descriptions presented by Russell as early as 1905 in his classic article "On Denoting." In Ramsey's words, the analysis Russell presented there formed a "paradigm of philosophy" (Ramsey 1931, 263). Definite descriptions are descriptions of the form of "the so-and-so" which apply to at most one individual; for example, "the oldest man in the world," or "the current president of Finland." Frege had not clearly distinguished between proper names and definite descriptions, but treated the latter in a way as a subcategory of simple individual names. This has certain undesirable consequences: If a definite description (e.g., "the current king of France") is not realized by any entity in the world, a sentence containing it seems to have no truth value. However, this is ill-suited for classic logic with "the law of the excluded middle"—the thesis that every meaningful (declarative) sentence is either true or false—to which both Frege and Russell were officially committed.[14]

Russell put forward a more sophisticated analysis and sought to show that sentences containing definite descriptions can be converted into sentences with the same meaning in which definite descriptions do not occur at all, i.e., that they can always be eliminated. Russell argued that a sentence with a definite description, such as "the current king of

[13] Korhonen 2013 is a rich and useful source on the earlier Russell.

[14] I do not intend to suggest that this is the route through which Russell actually ended up with his theory of descriptions. The complicated story can be found in Makin 2000.

France is bald," actually means the same as the conjunction of the following three sentences:

(1) There is at least one entity that is the current king of France;

(2) There is no more than one entity who is the present king of France; and

(3) any object that is the current king of France is also bald.[15]

This much more complex sentence is thus, according to Russell, an analysis of the meaning of the original sentence in written out form and reveals its true logical form. Since the sentence (1) is false, the whole long combined sentence (1)–(3) and thus also (according to Russell) the original sentence "the current king of France is bald," which means the same, are also false. This is a clear improvement over Frege's simpler intuitive approach, e.g., in that it is compatible with the law of the excluded middle. (However, all the tools of Russell's analysis were already included in Frege's logic. Frege just never figured to take the decisive step.)

Russell clearly came up with a new kind of idea of analysis here: In the earlier thought of Russell and Moore, "analysis" had meant the *metaphysical analysis* of reality into its fundamental building blocks—literally division into parts. The analysis now envisaged, on the other hand, focuses on language and sentences, and the sentence to be analyzed is transformed sometimes into a very different form. This sort of analysis has been called "transformative analysis" (see Beaney 2002, 2007b).[16]

In fact, such an analysis already appears in Frege's work in his analysis of the concept of number, although it did not yet at that time become a more general model for doing philosophy. Namely, Frege and also early Russell put forward an analysis of the concept of the natural number as follows: they suggested that a sentence involving a certain natural number, say 4, for example, "Jupiter has four moons," should be ana-

---

[15] More formally, the sentence has the logical form:
$(\exists x)(K(x) \wedge (\forall y)(K(y) \rightarrow x{=}y) \wedge B(x))$.

[16] The same thing is sometimes (including Beaney himself) alternatively called "interpretive analysis," and also "logical analysis"; but I personally find "transformative analysis" a more descriptive and apt label.

lyzed as "the concept *moon of Jupiter* has four instances." In other words, the sentence does not actually predicate the property *has four moons* of Jupiter, but rather predicates a (second order) property *has four instances* of the (first-order) concept *moon of Jupiter*. The purpose of such an analysis is to reveal the "real" logical form of the sentence to be analyzed.

One might perhaps argue that the idea of the unsatisfactory quality of natural language from the point of view of logic, and also the idea of transformative analysis, is already contained in the rejection of the subject-predicate structure of natural language and replacement of it by the function-argument structure (a starting-point of Frege's new logic), and in particular in the thesis of the ambiguity of "is" which was at the heart of the new logic of Frege and Russell.[17] Namely, Frege and Russell suggested that the following different meanings could be distinguished in the verb "is":

Meaning:	Example:
1. The *is* of identity	Saul is Paul
2. The *is* of predication (copula)	Paul is an apostle
3. The *is* of class inclusion	A vole is a mammal
4. The *is* of existence	God is

According to Frege and Russell, these are logically utterly different things, even though natural language uses the same verb "is" for all of them. This thesis is built into the whole new logic developed by Frege and Russell, for in it all the above things are expressed in quite different notation—in contrast to the ambiguous natural language:[18]

1. $a = b$ 2. $P(a)$ 3. $(\forall x)\,[(P(x) \rightarrow Q(x)]$ 4. $(\exists x)\,(x = a)$

In all these cases (i.e., the analysis of the concept of number, the ambiguity of "is," and in Russell's case, also the analysis of definite descriptions), it seems that the surface form of or-

[17] See Haaparanta 1985, 1986 (these focus on Frege).

[18] I am using throughout this paper the familiar notation now common in logic; not Frege's quite idiosyncratic two-dimensional notation nor Russell's notation which he adopted from Peano (the latter is closer to the modern one).

dinary language is unreliable and can mislead us and result in confusions, and only an analysis in terms of the constructed ideal language reveals the true logical form of the sentences at stake and dissolves confusions.

In 1902, Russell found a contradiction in Frege's grand system of logic, what is now called "Russell's paradox." Russell himself sought to develop a paradox-free general logic, which he began to call "the theory of types" (Russell 1908; the basic idea appears already in Russell 1903b).[19] The comprehensive presentation of the system was the ponderous three-part *Principia Mathematica* co-written with Whitehead (Russell & Whitehead 1910–1913). In the theory of types, predicates have their own restricted ranges of significance, properties have their own "types," and known paradoxes are ungrammatical and hence impossible to formulate. Its language clearly distinguishes, at the grammatical level, first-order predicates related to the properties of individual objects, predicates related to the properties of such properties (second-order properties), etc. In Russell's view, natural language is deficient in this case too, as it does not distinguish between them but makes the properties of different orders appear to be on an equal footing, which then results in contradictions.

Inspired by these phenomena, it was quite natural to think that perhaps at least some of the eternal problems of philosophy that seemed unsolvable would be revealed in logical analysis to be ungrammatical and thus meaningless (e.g., Carnap; see below). This vision has played an important role in making new formal logic such an integral part of contemporary philosophy. Russell's idea of the ranges of significance of the concept also influenced Ordinary Language Philosophy, where philosophical problems were sometimes interpreted to result from "category mistakes" (esp. Gilbert Ryle 1938, 1949).[20]

In more detail and explicitly, Russell described his own conception of a logical ideal language in his lectures on the

[19] See Urquhart 2006 for an overview.

[20] Carnap's famous 1931 paper on overcoming metaphysics (see below) may perhaps have also been an influence; it discusses (briefly) examples very similar to those of Ryle, under the label "type confusions" with an explicit reference to the Russellian theory of types.

philosophy of logical atomism from 1918 (they already show clearly Wittgenstein's influence):

> In a logically perfect language the words in a proposition would correspond one by one with the components of the corresponding fact, with the exception of such words as "or", "not", "if", "then", which have a different function. In a logically perfect language, there will be one word and no more for every simple object, and everything that is not simple will be expressed by a combination of words, by a combination derived, of course, from the words for the simple things that enter in, one word for each simple component. A language of that sort will be completely analytic, and will show at a glance the logical structure of the facts asserted or denied.
>
> The language which is set forth in *Principia Mathematica* is intended to be a language of that sort. It is a language which has only syntax and no vocabulary whatsoever. Barring the omission of a vocabulary I maintain that it is quite a nice language. It aims at being the sort of a language that, if you add a vocabulary, would be a logically perfect language. Actual languages are not logically perfect in this sense, and they cannot possibly be, if they are to serve the purposes of daily life. A logically perfect language, if it could be constructed, would not only be intolerably prolix, but, as regards its vocabulary, would be very largely private to one speaker. (Russell 1918, 197–198)

When Russell says here that the language is analytic, it does not mean that the sentences in the language are analytically true but that all the sentences in the language are fully, completely analyzed sentences. The structure of such a language thus undistortedly reflects the metaphysical structure of the world. The distinctions and categories of language are thus also the distinctions and categories of the world, metaphysical categories. The structure of the world can be read directly from the structure of the ideal language.

According to Russell, in a logically perfect language, communication from one speaker to another is impossible, except for matters of logic. Since ordinary language is not logically perfect, a philosopher who wants to find out the true logical form of a statement must analyze the statement and transform it into some, perhaps very different, sentence of a logically perfect language. Apparently, Russell also assumed

that the fully analyzed form corresponds to the structure of a thought expressed by the unanalyzed sentence in ordinary language and corresponds to something psychologically real. For Russell, thought is more primary than linguistic expression, and ordinary language often only imperfectly expresses the thought.

In his introduction to the English edition of Wittgenstein's *Tractatus* (see below), Russell wrote:

> A logically perfect language has rules of syntax which prevent nonsense, and has single symbols which always have a definite and unique meaning. Mr Wittgenstein is concerned with the conditions for a logically perfect language – not that any language is logically perfect, or that we believe ourselves capable, here and now, of constructing a logically perfect language, but that the whole function of language is to have meaning, and it only fulfills this function in proportion as it approaches to the ideal language which we postulate. ... The first requisite of an ideal language would be that there should be one name for every simple, and never the same name for two different simples. (Russell 1922, 8–9)

From his pivotal article on definite descriptions from 1905 onwards, Russell considered "the principle of acquaintance," as he called it, to be the very central guiding rule for constructing the ideal language:

> Every proposition which we can understand must be composed wholly of constituents with which we are acquainted.[21]

According to Russell, we can all be acquainted with the same abstract objects. Therefore, we can communicate about logic and mathematics, even if each of us spoke a logically perfect language. In contrast, we are not acquainted with physical objects or other minds. For example, in 1912, Russell thought that we can only be acquainted with the following: sense data, inner data, and memories of such things—and "perhaps" of the Self. Interpreted in this way, the principle of acquaintance imposes very severe conditions on the nature of fully analyzed sentences, and thus also on a logically perfect lan-

[21] This formulation is from Russell 1912; there is a slight variation in formulations in different works of Russell.

guage: The sentences of the language involve only logical constants, abstract objects or universals, and data from internal and external senses and memories of such—besides abstract objects, these are data of no more than one subject. I know my own sense data, you know yours, etc. Russell's logically perfect language is thus essentially a language of one person. With the sole exception of abstract objects, the sentences of my logically perfect language contain only words that refer solely to objects that no one else but I know and can be acquainted with. The logically perfect language outlined by Russell was indeed the paradigm of private language criticized by Wittgenstein in his later philosophy (cf. Hylton 2007).

After finishing *Principia*, the exhausted Russell moved in the 1910s from logic and the philosophy of mathematics to work mainly in epistemology. In addition to the principle of acquaintance, he was now guided by Occam's razor and his "supreme maxim in scientific philosophizing," formulated in 1914: "Whenever possible, inferred entities must be replaced by logical constructions." Still, in 1912, Russell had regarded the material objects of everyday life—i.e., rocks and trees, cats and dogs, tables and chairs—as inferred entities which explain and cause sense data. However, this opened the door to skepticism, and Russell did not tolerate the situation for long. He began to think that material objects should be given a treatment similar to the one he had given to numbers: words that seem to refer to material objects should be defined in terms of words that refer to things with which we are acquainted. From 1914 onwards, Russell thought that material objects were mere "logical constructions" out of sense data. By 1914, Russell's conception of philosophy also seemed to become more austere. He now wrote: "Every philosophical problem, when it is subjected to the necessary analysis and purification, is found to be not really philosophical at all, or else to be, in the sense in which we are using the word, logical" (1914, 42). Although this was not yet quite Wittgenstein's full-blown radical view of philosophy (see below), it certainly looks like a move in that direction, paving the way for it.

Be that as it may, Russell's matured position has in fact truly radical consequences for analysis and a logically perfect language: A fully analyzed form of even simple everyday

sentences would be thus astronomically complex and practically humanly unattainable—as would be a logically perfect language. A complete analysis would only be possible in logic and mathematics. One might think that a philosopher benefits already from partial, incomplete analyses—that we can at least get closer to a fully analyzed sentence. However, given the large-scale transformations the sentences go through in such an analysis, there is no good reason to assume that each intermediate step would be closer in logical form to the actual logical form than the previous one. The conclusion is quite discouraging for philosophy (cf. Hylton 2007).

5. Wittgenstein and the Logico-Philosophical Treatise

In 1911, Ludwig Wittgenstein (1889–1951), a young Austrian student of engineering who had become interested in philosophical problems in mathematics, sought to become a student of Russell at Cambridge. Soon the ingenious student began to influence his already famous teacher. Wittgenstein's early philosophy culminated in a small book with a downright cult reputation, *Tractatus Logico-Philosophicus* (1921).[22] The idea of an ideal logical language played an important role also in Wittgenstein's thought during that period.[23] It was familiar to him from Frege's work and from Russell through both his writings and their personal conversations. Right away in the introduction to *Tractatus*, he states that philosophical problems are based on a "misunderstanding of the logic of our language." Later in the book, he argues that "[m]ost questions and propositions of the philosophers result from the fact that we do not understand the logic of our language" (4.003).[24]

---

[22] For illuminating discussions on the aims and arguments of this short but difficult tractate, see Ricketts 1996, Kremer 2013.

[23] I shall simply put aside the difficult question of the correct interpretation of *Tractatus* as a whole, and how the relation of its quite skeptical conclusions and more constructive parts should be understood; I will only summarize how the theme of this paper appears in *Tractatus*. I leave it for Wittgenstein scholars to dispute whether and in what sense those statements are in the end themselves nonsensical and devoid of meaning, as the final mysticist paragraphs of the book suggest.

[24] In full:

According to early Wittgenstein too, ordinary language is a source of confusion: "In the language of everyday life it very often happens that the same word signifies in two different ways—and therefore belongs to two different symbols—or that two words, which signify in different ways, are apparently applied in the same way in the proposition." (3.323) Wittgenstein immediately gives, as an example, the ambiguity of the expression "is" emphasized by Frege and Russell: "Thus the word 'is' appears as the copula, as the sign of equality, and as the expression of existence"; and "In the proposition 'Green is green'—where the first word is a proper name as the last an adjective—these words have not merely different meanings but they are different symbols." And this, in Wittgenstein's mind, has a philosophical significance: "Thus there easily arise the most fundamental confusions (of which the whole of philosophy is full)." (3.324)

Therefore, according to Wittgenstein, an ideal language is needed: "In order to avoid these errors, we must employ a symbolism which excludes them, by not applying the same sign in different symbols and by not applying signs in the same way which signify in different ways. A symbolism, that is to say, which obeys the rules of logical grammar—of logical syntax." (3.325) Wittgenstein adds that "[t]he logical symbolism of Frege and Russell is such a language, which, however, does still not exclude all errors."

Indeed, Wittgenstein arrives at a very radical conception of the nature of philosophy: "All philosophy is 'Critique of language' … Russell's merit is to have shown that the apparent logical form of the proposition need not be its real form." (4.0031) Here Wittgenstein is gesturing, of course, toward Russell's analysis of sentences containing definite descrip-

---

"4.003 Most propositions and questions, that have been written about philosophical matters, are not false, but senseless. We cannot, therefore, answer questions of this kind at all, but only state their senselessness. Most questions and propositions of the philosophers result from the fact that we do not understand the logic of our language.

(They are of the same kind as the question whether the Good is more or less identical than the Beautiful.)

And so it is not to be wondered at that the deepest problems are really no problems."

tions.[25] This statement of Wittgenstein incidentally marked the start of the whole orthodox language-focused analytic philosophy.

Although he did not much develop an ideal language more formally, Wittgenstein was not an unoriginal follower of Frege and Russell either; his philosophical interpretation of an ideal language differed in certain important respects from theirs. For example, Frege and (earlier) Russell viewed also logical constants (such as "¬" and "∨") as names that—in order to be meaningful—must have some kind of abstract logical objects as their referents. According to Wittgenstein, in contrast, logical constants do not denote anything: they are not the names of any objects or complexes of objects. Wittgenstein, for example, suggested that the sentences "*P*" and "¬¬*P*" say the same thing or have the same content—if "¬" were a name, however, they would have radically different meanings.

However, the key difference between them in relation to natural languages and artificial ideal languages is the following: Frege and Russell thought that natural languages are logically flawed because they contain vague words and misrepresent the object of logic. Wittgenstein, in contrast, argued that "All propositions of our colloquial language are actually, just as they are, logically completely in order" (5.5564). The sentences of natural language are not, according to him, less logically correct or more logically confused than the sentences formed in the ideal languages of Frege or Russell. (Of course, the correct logical form of sentences is easier to see in an ideal language.) For Wittgenstein, logic is a prerequisite for all meaningfulness. Thus, nothing like illogical language can simply exist. If a sign has a sense at all, it must be logically in order. Thus, natural languages only seem to be logically flawed, according to Wittgenstein.[26]

[25] For much more about Russell's "merit" here, see Kremer 2012.

[26] For recent discussions on Wittgenstein and the limits of language, see the various essays in Appelqvist 2020.

6. Carnap: From rational reconstruction to explication

Rudolf Carnap (1891–1970), a German-born logician-philosopher and one of the central logical positivists of the Vienna Circle, has often been considered the paradigmatic representative of Ideal Language Philosophy.[27] Carnap continued the tradition of Frege and Russell and believed in the superiority of artificial formal languages in conducting philosophical research and used them essentially in his own philosophical investigations. He had attended Frege's lectures in 1914 and was greatly impressed by Russell's logical work. He was also deeply influenced by Wittgenstein's *Tractatus*. Carnap wanted to replace natural language even in everyday communication with a better artificial substitute: he was an active advocate of Esperanto.

Early Carnap continued in many ways from where Russell had left off. However, as a radical empiricist, he did not allow abstract objects or "the inner sense" that would enable acquaintance with such, as Russell had done. As a logical positivist, he also took a very negative view of all metaphysics and thus did not think that an ideal language would reflect the structure of any external reality. In this sense, he did not believe in any logically perfect language in the sense that Russell did. However, Carnap initially thought that Russell's logical system provided more or less the only possible and absolutely correct language of logic.

In his early classic work *Der logische Aufbau der Welt* ("The Logical Structure of the World") (1928), Carnap refers at the outset to Russell's supreme maxim in scientific philosophizing: whenever possible, inferred entities must be replaced by logical constructions. Indeed, in this work Carnap seeks to carry through in detail the program outlined by Russell of the logical construction of our knowledge of physical reality with mere sense data as a starting point (few believe that his attempt succeeded).[28] Carnap also refers at the beginning of this work to Leibniz's idea of an ideal language. At the time, he called his project "rational reconstruction"—i.e., he aimed

[27] Leitgeb & Carus 2020 gives a rather encompassing review of Carnap's thought.

[28] There are, however, some substantive differences between Carnap's approach and that of Russell; see, e.g., Beaney 2004.

to clarify old concepts by giving them new, more precise definitions.

Among other things, his article on the rejection of metaphysics (Carnap 1931)—famous for its critique of Heidegger, albeit it represents only a brief interphase in Carnap's thought—clearly shows his sour attitude toward natural language. According to Carnap, metaphysical statements are devoid of meaning either because they contain meaningless words, or because they combine meaningful words in a way that violates the logical syntax, i.e., the rules of sentence formation. Carnap called the latter type of apparent statements "pseudo-statements": They look like statements, but in reality do not state anything and do not express true or false statements. Carnap writes: "The fact that natural languages allow the formation of meaningless sequences of words without violating the rules of grammar, indicates that grammatical syntax is, from a logical point of view, inadequate. If grammatical syntax corresponded exactly to logical syntax, pseudo-statements could not arise." (Carnap 1931, 68) "It follows," Carnap continues, that "metaphysics could not even be expressed in a logically constructed language. This is the great philosophical importance of the task, which at present occupies the logicians, of building a logical syntax." (Ibid.)

According to him, "perhaps the majority" of the logical errors that underlie pseudo-statements are based on the ambiguity of the expression "to be" (or "is") in natural language (an apparent gesture toward Frege and Russell). Another very common violation of the correct logical syntax is, according to Carnap, "type confusions" of concepts, i.e., a natural language sentence which conflicts the sentence-formation rules and meaningful ranges of significance of predicates in Russell's theory of types.

Soon, however, Carnap abandoned the whole idea of one correct logical language[29] and adopted his famous Principle of Tolerance:

> Principle of Tolerance: It is not our business to set up prohibitions, but to arrive at conclusions. [Carnap 1934, §17]

[29] Carnap's quite sudden and radical change of view is tracked in Awodey & Carus 2007.

> In logic, there are no morals. Everyone is at liberty to build his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments. [Carnap 1934, §17]

Thus, a wide variety of alternative logical languages were now equally permitted. However, they must be presented with precise grammatical rules. Therefore, they cannot be natural languages but must be alternative artificial, formal languages. From now on, the choice of language system for Carnap was a pragmatic question: one language may be more useful for one purpose, another for another purpose. However, the choice of a language is not meaningfully a question of right and wrong, or true and false. In the 1930s, Carnap's philosophical inquiries were restricted to syntax, and for him all autonomous and legitimate philosophy that did not reduce to empirical sciences—such as psychology—and was not meaningless metaphysics was limited to studying the logical syntax of the language of science: "Philosophy is to be replaced by the logic of science" (Carnap 1934).

However, the syntactic perspective soon proved too restrictive, and in particular under the influence of Tarski (see below), Carnap expanded his conceptual framework to include the semantics of language, i.e., the meaning relations of language to the world and its objects. However, Carnap's analyses still focused on artificial formal languages. Indeed, after his "semantic turn," Carnap made a distinction between pure and descriptive semantics (see Carnap 1942, 11–15). Descriptive semantics is concerned with historically given natural languages, such as German, and is based on empirical investigation. Pure semantics, on the other hand, is an analysis of semantical systems with artificial languages which are stipulatively defined. It is entirely analytic and without factual content. "Here we lay down definitions for certain concepts, usually in the form of rules, and study the analytic consequences of these definitions. In choosing the rules we are entirely free," he explains (Carnap 1942, 13). Philosophy then, according to Carnap, must confine itself to pure semantics. For Carnap, pure and descriptive semantics seem to be independent and autonomous projects.

At this point, a new kind of conception of analysis began to emerge more and more clearly in Carnap's thought: Instead of rational reconstruction, he started to talk about "explication." He borrowed the term from Husserl, even though these two philosophers meant somewhat different things with it. Carnap's explication relies essentially on artificial formal languages. Explication is clarification or "refining" of meaning. The criterion for the goodness of its results may be their ability to clarify the meaning of the old term in a way that highlights one of its key "meanings" (ambiguous terms) or covers "clear cases" in the extensions of the original inaccurate term, and creates and clarifies links with other scientific concepts. They are expected to have not only "usability" but also "theoretical fertility" and "systematic strength."[30]

In his groundbreaking work on the semantics of intensional logic and possible worlds semantics, *Meaning and Necessity* (Carnap 1947), Carnap says he seeks to clarify the concept of *meaning*. At the same time, he describes the idea of explication as follows:

> The task of making more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather if replacing it by a newly constructed, more exact concept, belongs among the most important tasks of logical analysis and logical construction. We call this the task of explicating, or of giving an explication for the earlier concept: this earlier concept, or sometimes the term used for it, is called the *explicandum*; and the new concept, or its term, is called an *explicatum* of the old one. (Carnap 1947, 7–8)

Although Carnap's actual project here is an explication of the concept of *meaning*, he gives as an example of explication the analysis of the concept of *number* by Frege and Russell:

> Thus, for instance, Frege and, later, Russell, took as an *explicatum* the term "two" in the not quite exact meaning in which it is used in everyday life and in applied mathematics; they proposed as an *explicandum* for it an exactly defined concept, namely, the class of pair classes. (Carnap 1947, 8)

---

[30] The fullest presentation of his conception of explication is in Carnap 1950. Beaney 2004 includes a thorough discussion of it.

Carnap mentions, as other examples of explication, Russell's analysis of definite descriptions and Tarski's semantic analysis of the concept of truth (see below). He adds:

> Generally speaking, it is not required that an *explicatum* have, as nearly as possible, the same meaning as the *explicandum*; it should, however, correspond to the *explicandum* in such a way that it can be used instead of the latter. (Carnap 1947, 8)

Were the analyses of Frege and Russell then cases of explication in Carnap's sense, or just analyses of already existing meanings? On the one hand, their own explicit comments may suggest that the latter is the case. On the other hand, their critical views concerning natural languages make it somewhat difficult to understand how it could be. Therefore, Carnap may be on the right track when he is suggesting that the concept of explication he presents describes better what they were actually doing: Frege and Russell may not have sufficiently distinguished between the two.[31]

Later, Carnap puts forward four requirements for a good *explicatum*: 1) it is to be similar to the *explicandum* in such a way that it can be used in most cases in which the *explicandum* has so far been used; 2) the rules of its use are to be given in an exact form, in conjunction with other scientific concepts; 3) it is to be a fruitful concept, i.e., useful for the formulation of many universal statements; and 4) it should be as simple as possible, given the more important requirements (1)–(3) (Carnap 1950, 7).

Carnap advocated to the end the fundamental thesis—inherited from Wittgenstein—that philosophy is primarily an activity of clarifying language, and makes no claims and presents no theories. More specifically, it came to mean to him that all legitimate philosophy amounts to the activity of explicating concepts by means of artificial formal languages.

7. Tarski and the inconsistency of natural language

The Polish logician Alfred Tarski (1901–1983) can also be naturally viewed as a representative of Ideal Language Philoso-

[31] In fact, in a relatively late lecture (Frege 1914), Frege sketches a notion of definition which is not that different from Carnap's idea. What is more, Carnap attended that lecture of Frege in Jena. See Beaney 2004.

phy. He was the father of logical semantics and one of the most significant logicians of our time. He is known in philosophy especially for his theory of truth based on the tools of formal logic (Tarski 1933/1935, 1944). Tarski's influence was also essential when Carnap turned from the syntactic approach to the semantic one in the late 1930s. Tarski was primarily a logician, and unlike Wittgenstein or Carnap, he did not put forward any general theses on the task and the nature of philosophy. In practice, however, his work on truth has been one of the best-known examples of the ideal language tradition.[32]

Tarski contended that truth can only be defined in formal languages and only one at a time. Natural languages, he suggested, are "semantically closed," meaning they can talk about their own truth and other semantical properties. This in turn leads to many paradoxes and contradictions, e.g., "the liar paradox." Therefore, according to Tarski, the concept of truth can be unequivocally defined only for languages which are "semantically open" and which have precisely defined rules of grammar. (See Tarski 1944.) Tarski's great influence on Carnap may suggest that their philosophical attitudes are also more or less the same. However, there are in fact some interesting differences between them.

To begin with, for Tarski, the "formal languages" whose truth is under consideration must always be interpreted languages, not purely formal, as he repeatedly emphasized:

> It remains perhaps to add that we are not interested here in 'formal' languages and sciences in one special sense of the word 'formal', namely sciences to the signs and expressions of which no meaning is attached. For such sciences the problem here discussed has no relevance, it is not even meaningful. We shall always ascribe quite concrete and, for us, intelligible meanings to the signs which occur in the languages we shall consider. (Tarski 1933/1935, 166–67)

Nor was this just an occasional philosophical opinion for Tarski; it was quite an integral part of Tarski's whole ap-

[32] Gómez-Torrente 2019 gives a good overview of Tarski's work. For more about philosophical aspects of Tarski's thought, see, e.g., Woleński 1993, Mancosu 2009, Patterson 2012.

proach to truth that the meanings of the object language must be given and fixed. Only in this way can the definition of truth (applied to sentences) make any sense at all:

> For several reasons it appears most convenient to apply the term "true" to sentences, and we shall follow this course.[footnote omitted] Consequently, we must always relate the notion of truth, like that of a sentence, to a specific language; for it is obvious that the same expression which is a true sentence in one language can be false or meaningless in another. (Tarski 1944, 342)

> We shall also have to specify the language whose sentences we are concerned with; this is necessary if only for the reason that a string of sounds or signs, which is a true or a false sentence but at any rate meaningful sentence in one language, may be a meaningless expression in another. (Tarski 1969, 64)

> . . . the concept of truth essentially depends, as regards both extension and content, upon the language to which it is applied. We can only meaningfully say of an expression that it is true or not if we treat this expression as a part of a concrete language. As soon as the discussion concerns more than one language the expression "true sentence" ceases to be unambiguous. If we are to avoid this ambiguity we must replace it by the relative term "a true sentence with respect to the given language." (Tarski 1933/1935, 263)

Tarski made a distinction, which resembled Carnap's distinction of descriptive and pure semantics, between descriptive semantics and theoretical semantics (Tarski 1944). By "descriptive semantics," he refers to the totality of the study of semantic relations in natural languages. "Theoretical semantics" apparently means to Tarski the kind of research he does himself.

It is true that Tarski constantly stressed that natural languages drift into semantic paradoxes, and that truth can be unequivocally defined only for formal languages. This has led many to assume that Tarski, like Carnap, would have liked to limit his "theoretical semantics" to artificial formal languages only—that it could not be applied at all to real-life natural languages. In the case of Tarski, however, the matter is more complicated. We have already seen above that, for-

mal or not, the languages in question must, for Tarski, be "concrete" and already interpreted, that is to say, already provided with "concrete" meanings. This alone makes them quite different from artificial formal languages in the conventional sense.

Tarski also thought that his formal semantic tools could be applied to the limited languages of various special sciences, such as chemistry, so long as they did not contain semantic vocabulary. Furthermore, Tarski suggested that theoretical semantics is, after all, applicable to natural languages, albeit "only with certain approximation" (Tarski 1944, 365). Namely: "the approximation consists in replacing a natural language (or a portion of it in which we are interested) by one whose structure is exactly specified, and which diverges from the given language 'as little as possible'" (Tarski 1944, 347). Tarski also writes that "[t]he results obtained for formalized language also have a certain validity for colloquial language ... if we translate into colloquial language any definition of a true sentence which has been constructed for some formalized language, we obtain a fragmentary definition of truth which embraces a wider or narrower category of sentences" (Tarski 1933/1935, 165). In fact, at one point, Tarski stressed that when he used the term "formal language," he did not "have in mind anything essentially opposed to natural languages"; he continues, "[on]n the contrary, the only formal languages that seem to have real interest are fragments of natural languages (fragments provided with complete vocabularies and precise syntactic rules) or ones that can at least be sufficiently translated into natural languages" (Tarski 1969, 68). Tarski's attitude toward natural language was thus in fact somewhat less hostile than that of Carnap.

8. Afterword

The new formal logic developed by Frege and Russell, or rather the first-order logic contained in it as a proper part, has become an established and familiar tool of philosophers. Carnap's work has been an important point of departure in both the philosophy of language and the philosophy of science, and Tarski's formal theory of truth is a mandatory basic theory in all philosophical theorizing of truth. The tools of

formal logic continue to play a central role in philosophy, especially in the philosophy of mathematics and the philosophy of science, and in some respects also in the philosophy of language, and even in metaphysics.

However, few still believe, like Carnap did, that all that philosophy can legitimately do is engage in clarificatory activity focusing on language. Today, philosophers incontinently present arguments with conclusions and advocate philosophical theories. At the same time, the idea that formalized languages would play such a central role in all philosophical activity as suggested by Carnap and some others has become quite rare. The attitude of philosophers toward the tools of formal logic today is usually more pragmatic: they have their own fruitful applications, but often recourse to them is neither necessary nor useful. Many follow in practice Quine's humorous "maxim of shallow analysis": "where it doesn't itch don't scratch" (Quine 1960, 160).

It is perhaps natural to end the present overview with a quote from Saul Kripke (1940–2022), who is arguably one of the most important philosophers of our time. What is interesting here is that he has significantly followed in the footsteps of Carnap and Tarski in the study of intensional logic and the logical theory of truth and has been one of the most brilliant logicians in contemporary philosophy. According to him, the use of the tools of formal logic is sometimes useful in philosophy, but it must be informed by a sensitivity to the philosophical significance of the formalism and by a generous admixture of common sense. Kripke stated: "It should not be assumed that formalism can grind out philosophical results in a manner beyond the capacity of ordinary philosophical reasoning. There is no mathematical substitute for philosophy." (Kripke 1976, 416)[33]

Tampere University

[33] This paper is a somewhat revised and modified translation of my earlier Raatikainen 2013b (in Finnish). I am very grateful to Leila Haaparanta for her valuable comments on an earlier version of this paper.

References

Appelqvist, H. (ed.) (2020). *Wittgenstein and the Limits of Language*. New York: Routledge.

Awodey, S., & A. W. Carus (2007). "Carnap's Dream: Gödel, Wittgenstein, and 'Logical syntax.'" *Synthese* 159, 23–45.

Beaney, M. (1996). *Frege: Making Sense*. London: Duckworth.

Beaney, M. (2002). "Decompositions and Transformations: Conceptions of Analysis in the Early Analytic and Phenomenological Traditions." *Southern Journal of Philosophy*, 40, Supp. Vol., 53–99.

Beaney, M. (2004). "Carnap's Conception of Explication: From Frege to Husserl?" In S. Awodey & C. Klein (eds.), *Carnap Brought Home: The View from Jena*. Chicago: Open Court, 117–50.

Beaney, M. (ed.) (2007a). *The Analytic Turn: Analysis in Early Analytic Philosophy and Phenomenology*. London: Routledge.

Beaney, M. (2007b). "The Analytic Turn in Early Twentieth-century Philosophy." In Beaney 2007a, 1–30.

Beaney, M. (ed.) (2013). *The Oxford Handbook of The History of Analytic Philosophy*. Oxford: Oxford University Press.

Beaney, M. (2016). "The Analytic Revolution." *Royal Institute of Philosophy Supplement* 78, 227–249.

Carnap, R. (1928). *Der Logische Aufbau der Welt.* Leipzig: Felix Meiner Verlag.

Carnap, R. (1931). "Überwindung der Metaphysik durch logische Analyse der Sprache." *Erkenntnis* 2, 220–241. English translation: "The Elimination of Metaphysics Through Logical Analysis of Language." In A.J. Ayer (ed.), *Logical Positivism*. New York: The Free Press, 1959, 60–81. (Page references are to the translation).

Carnap, R. (1934). *Logische Syntax der Sprache.* Vienna: Julius Springer. English translation: *Logical Syntax of Language*. London: Routledge, 1937. (Page references are to the translation).

Carnap, R. (1942). *Introduction to Semantics*. Cambridge, MA: Harvard University Press.

Carnap, R. (1947). *Meaning and Necessity.* Chicago: University of Chicago Press.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Frege, G. (1879). *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Louis Nebert. English translation of the Preface and Part I: in Frege 1997, 47–78. (Page references are to the translation).

Frege, G. (1882). "Über die wissenschaftliche Berechtigung einer Begriffsschrift." *Zeitschrift für Philosophie und philosophische Kritik* 81, 48–56. English translation: "On the Scientific Justification of a Begriffsschrift." In Frege 1972, 83–89. (Page references are to the translation).

Frege, G. (1914). "Logic in Mathematics." In G. Frege, *Posthumous Writings*, ed. by H. Hermes et al. Oxford: Blackwell, 1979, 203–50.

Frege, G. (1972). *Conceptual Notation and Related Articles*, ed. by T. W. Bynum. London: Oxford University Press.

Frege, G. (1997). *The Frege Reader*. ed. by M. Beaney. Oxford: Basil Blackwell.

Gabriel, G. (2013). "Frege and the German Background to Analytic Philosophy." In Beaney 2013, 280–97.

Gómez-Torrente, M. (2019). "Alfred Tarski." In *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/ entries/tarski/>.

Haaparanta, L. (1985). *Frege's Doctrine of Being.* Acta Philosophica Fennica Vol. 39. Helsinki: Suomen Filosofinen Yhdistys.

Haaparanta, L. (1986). "On Frege's Concept of Being." In S. Knuuttila & J. Hintikka (eds.), *The Logic of Being*. Dordrecht: Reidel, 269–290.

Hylton, P. (2007). "'On Denoting' and the Idea of a Logically Perfect Language." In Beaney 2007a, 91–106.

Hylton, P. (2013). "Ideas of a Logically Perfect Language in Analytic Philosophy." In Beaney 2013, 906–925.

Korhonen, A. (2013). *Logic as Universal Science: Russell's Early Logicism and its Philosophical Context*. London: Palgrave-Macmillan.

Kremer, M. (2012). "Russell's Merit—The Obvious Interpretation." In J. L. Zalabardo (ed.), *Wittgenstein's Early Philosophy*. Oxford: Oxford University Press, 195–240.

Kremer, M. (2013). "The Whole Meaning of a Book of Nonsense: Introducing Wittgenstein's *Tractatus*." In M. Beaney 2013, 451–585.

Kripke, S. (1976). "Is There a Problem about Substitutional Quantification?" In G. Evans & J. McDowell (eds.), *Truth and Meaning*. Oxford: Oxford University Press, 325–419.

Leitgeb, H. & A. Carus (2020). "Rudolf Carnap." In *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/ entries/carnap/>.

Lenzen, W. (2004). "Leibniz's Logic." In D. M. Gabbay & J. Woods (eds.), *Handbook of the History of Logic, Volume 3. The Rise of Modern Logic: From Leibniz to Frege*. Amsterdam: Elsevier, 1–83.

Makin, Gideon (2000). *The Metaphysicians of Meaning: Russell and Frege on Sense and Denotation*. New York: Routledge.

Mancosu, P. (2009). "Tarski's Engagement with Philosophy." In S. Lapointe et al. (eds.), *The Golden Age of Polish Philosophy*. Dordrecht: Springer, 131–53.

Patterson, D. (2012). *Alfred Tarski: Philosophy of Language and Logic*. New York: Palgrave Macmillan.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Raatikainen, P. (2013a). "What Was Analytic Philosophy?" *Journal for the History of Analytical Philosophy* Vol 2, No 2, 11–27.

Raatikainen, P. (2013b). "Analyyttisen filosofian 'kova linja'." In J. Järvenkylä & I. Kortelainen (eds.), *Tavallisen kielen filosofia*. Helsinki: Gaudeamus, 178–202.

Ramsey, F. (1931). *The Foundations of Mathematics: and other logical essays*. London: Kegan Paul.

Ricketts, T. (1996). "Pictures, Logic, and the Limits of Sense in Wittgenstein's *Tractatus*." In H. D. Sluga & D. G. Stern (eds.), *The Cambridge Companion to Wittgenstein*. Cambridge: Cambridge University Press, 59–99.

Rorty, R. (1967). "Metaphilosophical Difficulties of Linguistic Philosophy." In R. Rorty (ed.), *The Linguistic Turn: Recent Essays in Philosophical Method*. Chicago and London: The University of Chicago Press, 1–41.

Russell, B. (1900). *A Critical Exposition of the Philosophy of Leibniz*. Cambridge: At the University Press.

Russell, B. (1903a). "Recent Work on the Philosophy of Leibniz." *Mind*, n.s., 12, 177–201.

Russell, B. (1903b). *The Principles of Mathematics*. Cambridge: At the University Press.

Russell, B. (1905). "On Denoting." *Mind* 56, 479–493.

Russell, B. (1908). "Mathematical Logic as Based on the Theory of Types." *American Journal of Mathematics* 30, 222–262.

Russell, B. (1912). *The Problems of Philosophy*. London: Williams and Norgate.

Russell, B. (1914). *Our Knowledge of the External World*. Chicago and London: The Open Court Publishing Company.

Russell, B. (1918). "The Philosophy of Logical Atomism." *Monist* 28, 495–527; 29, 32–63, 190–222, 345–380. Reprinted in: B. Russell, *Logic and Knowledge*, London: Allen and Unwin, 1956, 177–281. (Page references are to the reprint).

Russell, B. (1922). "Introduction." In L. Wittgenstein: *Tractatus Logico-Philosophicus*. London: Kegan Paul, 7–19.

Russell, B. (1944). "My Mental Development." In P. A. Schilpp (ed.), *The Philosophy of Bertrand Russell. Library of Living Philosophers.* Evanston, Illinois: Northwestern University Press, 1–20.

Russell, B. & A. N. Whitehead (1910–1913). *Principia Mathematica*, 3 vols. Cambridge: Cambridge University Press.

Ryle, G. (1938). "Categories." *Proceedings of the Aristotelian Society* 38, 189–206.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson University Library.

Sluga, H. D. (1980). *Gottlob Frege*. London: Routledge and Kegan Paul.

Tarski, A. (1933/35). *Pojecie prawdy w jezykach nauk dedukcyjnych*, Warsaw: Nakladem Towarzystwa Naukowego Warszawskiego (1933). In German, with a new postscript: "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Philosophica*, 1, 261–405 (1935). English translation: "The Concept of Truth in Formalized Languages." In A. Tarski: *Logic, Sematics, Metamathematics* (2nd edn.) ed. J. Corcoran. Indianapolis: Hackett, 1983, 152–278. (Page references are to the translation.)

Tarski, A. (1944). "The Semantic Conception of Truth and the Foundations of Semantics." *Philosophy and Phenomenological Research* 4, 341–376.

Tarski, A. (1969). "Truth and Proof." *Scientific American* 220, 63–77.

Trede, L. B. (1811) *Vorschläge zu einer notwendigen Sprachlehre*. Ohne Ortsangabe (Hamburg), 1811. 2. Auflage: Leipzig, 1816.

Trendelenburg, F. A. (1857). Über Leibnizens Entwurf einer allgemeinen Charakteristik. *Philosophische Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin*. Aus dem Jahr 1856, Berlin: Commission Dümmler, 36–69.

Urquhart, A. (2006). "The Theory of Types." In N. Griffin (ed.), *The Cambridge Companion to Bertrand Russell*. Cambridge: Cambridge University Press, 286–309.

Wittgenstein, L. (1921). "Logisch-Philosophische Abhandlung." In Wilhelm Ostwald (ed.), *Annalen der Naturphilosophie*, 14 (1921). English translation: *Tractatus Logico-Philosophicus.* London: Kegan Paul, 1922.

Woleński, J. (1993). "Tarski as a Philosopher." In F. Coniglione et al. (eds.), *Polish Scientific Philosophy: The Lvov-Warsaw School*. Amsterdam: Rodopi, 319–38.

Descartes on Language: How Signification Leads to Direct Reference[1]

JANI SINOKKI

1. Introduction

Many would find it trivially true to say that the *nature of ideas*, as Descartes conceives it, determines the kind of theory of intentionality he is committed to. However, to claim that the same is true of Descartes' *theory of reference* would almost surely not elicit such immediate concurrence.[2] Yet upon closer examination, both statements turn out to be just as trivially true. A reconstruction of Descartes' theory of intentionality answers the question of how ideas come to be about things so as to exhibit extra-mental things to the thinker—a topic of continued scholarly debate. Similarly, as far as his philosophical system goes, a reconstruction of Descartes' theory of reference explains how certain linguistic expressions, like names, connect to objects relevant for the truth-value of the sentences in which those expressions are used. In fact, Descartes' theory of intentionality—his view on the objective re-

[1] I am indebted especially to Joseph Almog for numerous discussions on issues related to the topic of this paper. I also thank Tapio Korte and Vili Lähteenmäki for their comments on a draft of this paper.

[2] "Theory of reference" is used here in the sense discussed, for example, in Raatikainen 2020, i.e., as a theory about language and (some important aspects of) linguistic representation. *Theory of intentionality*, on the other hand, is a view about the nature of the mind and mental representation. Therefore, and for the sake of clarity, in this paper "refer" is used only with regard to linguistic expressions, so that ideas or mental states don't refer but merely *have objects* (i.e., are *about*, or *of*, or *represent*, objects).

ality of ideas[3]—*is* what defines his views concerning linguistic reference.

In this paper, I will argue that Descartes is committed to a theory of *direct reference*.[4] According to this view, what a *singular term* brings to the "semantic value" of a sentence in which it is used is simply the object referred to.[5] *Reference*, in this sense, pertains to the relation between a linguistic item and the object that is relevant for determining the truth-value

---

[3] See CSM2 7, AT7 8; CSM2 29, AT7 41–42; CSM2 75, AT7 102. The abbreviation "CSM" refers to English translation of Descartes' works by Cottingham, Stootoff, and Murdoch (Descartes 1985; Descartes 1984) and is followed by the volume and page numbers, respectively. The third volume (Descartes 1991) includes Kenny as translator, and is abbreviated "CSMK" followed by page numbers. All English quotations are from CSM. Abbreviation "AT" refers to Adam and Tannery's edition of Descartes' works (Descartes 1904) and is followed by the volume and page numbers, respectively. For historical overviews of the doctrine of objective reality, see Read 1977; Nuchelmans 1983, chaps. 1–2; Normore 1986; Tachau 1988; and Ayers 1998. For a helpful comparison of objective reality in certain scholastics, as well as in Descartes, see Brown (2007, 139–43). In that volume, see also Clemenson 2007, King 2007, and Tweedale 2007.

[4] For exposition of varieties of direct reference, see Almog (2014, chap. 2). In general, the term "direct reference" is intended to be synonymous with "non-denotational reference," i.e., reference as a relation between a linguistic expression and an object unmediated by "modes of presentation." However, beyond this negative definition things are complicated (for instance, Recanati 1993, xii, points out that the negative thesis does not mean that no "modes of presentation" are involved, only that they cannot be what determine reference.)

For different takes on direct reference, see Soames 1987, 50, and Kaplan 1989, 493. Kaplan makes use of Russellian propositions (it is this view which is mostly discussed below in relation to his views), most visible in his work on *de re* belief; see Kaplan 2013; see also 1989, 493–97; 2012. The nature of *de re* belief (see, e.g., Eaker 2004; Stalnaker 2009; Burge 2012; Kaplan 2013) is very important for the discussion in this paper, but I will not employ the terminology of *de re—de dicto*.

[5] Names, indexicals, and variables are paradigmatic directly referential terms. In this paper, I will mostly ignore variables. *Singular terms* refer to, denote, or designate particular things, while *general terms* apply to many things. At the end of the paper, I briefly discuss what direct reference amounts to.

of a sentence in which that linguistic item is used.[6] My discussion proceeds by dissecting the more fundamental relations that, according to Descartes, ground this semantic relation. My claim that Descartes is a direct referentialist might seem odd at first sight, for Descartes is also committed to the *theory of signification*. According to this age-old, and often disparaged, view, words receive their meaning by signifying *ideas* in the mind of the speaker. Critics from Mill onwards have understood any such mentalistic theory as either amounting to an assertion that ideas, instead of worldly objects, are the referents of names of objects, or viewed the view as leading to hopeless subjectivism in other ways.[7] However, a theory of signification is not a theory of reference, but rather amounts to the claim that "intentionality takes place at the level of ideas, not words."[8] Also, the threat of subjectivism clearly depends on how the nature of ideas is conceived by the accompanying account of thought. Importantly for my purposes, though a theory of signification is not a theory of reference, it will indeed produce one when combined with a theory of ideas. Interestingly, both the emerging theory of reference as well as the nature of signification relation will vary from philosopher to philosopher, possibly even drastically, depending on how they view the relation between thought and its objects.

My argument for viewing Descartes as a direct referentialist is as follows: For Descartes, ideas gain their in-

---

[6] It is important to note that Descartes occasionally uses "to refer" (*referrer*) with regards to ideas, in a sense very close to our contemporary one, in relation to what he calls *material falsity of ideas*: "For it often happens in the case of obscure and confused ideas—and the ideas of heat and cold fall into this category— that an idea is referred to something other than that of which it is in fact the idea" (CSM2 163; AT7 233). It is not immediately clear whether cases in which an idea is successfully "connected to" an object also count as cases of "referring," for the evidence is insufficient to properly assess Descartes' views about referring in the sense he uses it. Almog argues that Descartes' use of *referrer* signifies a mode of "going back" to the thing that has already penetrated into the mind (in private communication; see also Almog 2014, 23). His view is at least compatible with the one I present in this paper.

[7] Mill 2011, 15; Frege 1956; and Wittgenstein 2009, §§244–271.

[8] Ott 2008, 294.

tentionality from the *objective reality* contained in the ideas—objective reality being another mode of existence for the extra-mental objects themselves.[9] The conjunction of Descartes' theories of ideas and signification thus results in a view much like that once held by Bertrand Russell, a view according to which singular propositions are complexes that can contain worldly objects, like Mont Blanc, as their constituent parts.[10] Combined with signification, Descartes' view entails direct reference (unlike the Russellian conception of propositions, which only supports direct reference).

However, Descartes' view avoids certain problems that haunt Russellian direct reference, and can explain, for example, how co-referential names can have different "cognitive values" despite there being only one object involved. This is indeed possible, for as Margaret Wilson (1978, 90) has observed, there is a difference between the objective reality of an idea and its *representational character*.[11] While the objective reality is just the worldly object that is the content of the idea, its representational character functions like a Fregean "sense" in that it is a *mode of presentation* of its object. However, contrary to (some standard readings of)[12] Frege, this representa-

---

[9] See footnote 3 above.

[10] Russell held this view only briefly, prior to 1905. He expressed commitment to the view in his *Principles of Mathematics* (1903) and in a letter to Frege written in 1904, but by the time he wrote "On Denoting" (1905) he had already rejected the view. Kaplan (2012) elaborates on the neo-Russellian framework of singular propositions. For more on the nature of Russellian propositions, see also Wettstein 1986, 1990; and Almog 2012.

[11] This difference amounts to a distinction between the "level of objective content" and the "level of representation," which Kaplan (2012, 140) views as the touchstone of a direct referentialist theory. Almog (see esp. 2005) rejects this and argues that the postulation of any kind of "content" to explain this distinction is incompatible with direct reference. Bianchi (2007), in turn, argues that representations can be treated as "vehicles" as opposed to objective contents, thus creating a centrist position. The view I attribute to Descartes in this paper resembles that of Bianchi's.

[12] Dummett 1973, and famously also Kripke 1980, view Fregean senses (*Sinne*) as that which determines the reference of names. For a contrasting view, see Korte 2022. When referring to Frege's views in this paper, I mean only the received Dummett-Kripke reading of those views, at the peril of ignoring views that reflect more accurately those of Frege's actual views.

tional character does nothing to determine the object of the idea, for an idea's having a representational character already presupposes it objectively containing a thing.[13] Consequently, when the idea is signified, this mode of presentation does nothing to determine the referents of one's words.[14]

Because of its slightly programmatic nature, my paper probably should be supplied with more caveats than I can sensibly add here. Defending my reading of Descartes as a historically accurate interpretation requires a separate paper, or even several papers. In this paper I am content to point how my discussion here relates to some issues of general scholarly interest, such as *true and immutable natures* and *clear and distinct perception*, but I will not be able to elaborate on the matter due to space limitations. Similarly, my examination how the theory of reference I develop for Descartes properly relates to discussions in contemporary philosophy of language marks only a beginning.[15] For example, I will only begin to sketch how the puzzles about *empty names* or *informative identities* can be successfully solved in the view I propound. In the footnotes, I will present some additional connections between my discussion and these other debates. Before examining objective reality and representational char-

---

[13] The Fregean view of reference determination has been criticized extensively by Almog (see, e.g., 1985; 2005; 2008b; 2012; 2014). For a critique of Almog's criticism, see Bianchi 2007. Many, including Almog, see any kind of commitment to representations as squarely incompatible with the idea of direct reference (see also Capuano 2015). Thus, though Almog (2008a) has argued for a view about Descartes' ideas similar to that which I defend, he certainly would not be comfortable with my relaxed use of the notion of "representational character." However, as Kaplan (e.g., 2013, 29) has pointed out, the same problems that talk of representations plausibly raise would be raised by any other mediators, including Kripkean causal chains. Therefore, whether they are representations or not, the direct referentialist must accept the fact that reference nevertheless requires some kind of "vehicle." For discussion of such "vehicles," see Bianchi 2007.

[14] Thus, Descartes' view offers an alternative way to understand Kaplan's famous thesis "No mentation without representation!" (Kaplan 2012, 153; see also Almog 2005; Eaker 2004).

[15] I have, however, elaborated my view concerning the role of causation in grounding referential relation of names already in my earlier work (see Sinokki 2022).

acter of ideas, in sections 3 and 4 respectively, I will proceed by discussing Descartes' general views about language and thought, and the nature of signification, in the section 2 below.

2. Language and thought

Descartes writes directly about language very little. He never engages in anything resembling a theory of language or philosophical semantics. In fact, Descartes is surprisingly quiet about language when compared to some of his scholastic predecessors, his Cartesian followers at Port Royal, his commentator and critic Thomas Hobbes, or the paradigmatic signification theorist, John Locke.[16] Most of what Descartes says about language is in the context of skepticism about animal thought and is not, at first glance, very useful in understanding Descartes' general views about language. Nevertheless, it is a useful place for me to start my examination. It will quickly become evident that Descartes' views about linguistic meaning depend on his views about the nature of thought.

Descartes believes that non-human animals cannot think.[17] For Descartes, this is evinced beyond any doubt by the fact that even the most sophisticated animals can only mimic sounds at best, but cannot engage in genuinely meaningful speech or the meaningful use of signs.[18] In a letter to Marquess of Newcastle on 23 November 1646, Descartes famously argues that "the reason why animals do not speak as we do is not that they lack the organs but that they have no

---

[16] See Ott 2003, chap. 1. I rely heavily on Ott's discussion on signification. For my views on Locke's philosophy of language, see Sinokki 2011 (in Finnish).

[17] Notoriously, according to Descartes this entails that non-human animals also lack moral worth, though some commentators argue against this (esp. Cottingham 1978). See also Harrison 1992.

[18] Descartes also thought that meaningful conversation was the surest sign of the presence of intelligence, of a mind, be it in an animal or machine. His view is thus not too distinct from that of Turing's famous proposal (see Turing 1950; see González 2020; see also Cottingham 1997). However, questions about *detecting* a mind should not be confused with question about *having* (or being) a mind.

thoughts."[19] The connection between *genuine* language-use and thought is very important. As Descartes makes clear in the letter, non-human animals readily use various signs just as skillfully as humans to signify passions like hunger, fear, and joy. The point is that animals cannot speak due to their inability to attach the right kind of *semantic content* to their signs, and this is essential for genuine language use: "there has never been known an animal so perfect as to use a sign to make other animals understand something which bore no relation to its passions."[20] As genuine language-use is impossible for animals despite their ability to signify their passions with sounds, it is evident that the latter is not sufficient for the former.

Some commentators emphasize how features of human speech, like its unlimited productivity, ground Descartes' conclusion that language-use requires the presence of an immaterial mind.[21] This is true enough, for Descartes considers genuine language-use to be productive and "the only certain sign of thought hidden in a body."[22] However, it is important to underline that this is *not* because Descartes thinks that language is the only empirically observable manifestation of the otherwise hidden private and subjective realm of thought (as some later philosophers would have it). Almost the opposite is true in fact. Features of speech can act as guides to thinking precisely because Descartes thinks human language-use owes *all* its semantical features to thought. The nature of human language is also, in an important sense, *public*—but so is the nature of thought as well (this publicity of language and thought is a recurring topic in this paper).[23]

---

[19] CSMK 302–304; AT4 569–576.

[20] CSMK 303; AT4 575. Contrary to a common misconception, Descartes never denied that animals are capable of sensibility or communication: "all animals easily communicate to us, by voice or bodily movement, their natural impulses of anger, fear, hunger, and so on" (CSMK 366, AT5 278; see also Cottingham 1978; Harrison 1992).

[21] E.g., Cottingham 1997; Chomsky 1991.

[22] CSMK 366; AT5 278.

[23] This, of course, is in line with Descartes' widely documented general tendency towards reductivism and naturalism about meaning; see, e.g., Nolan 1997b; Alanen 2008.

Descartes seems committed to a view about the nature of language and the way words gain their semantic properties known as *theory of signification*: "whenever I express something in words, and understand what I am saying, this very fact makes it certain that I have an idea of what is signified by the words in question."[24] However, he never develops specific views about signification. Without a doubt he was very familiar with the theory, as it was regularly discussed in scholastic logic books.[25] In general, the view that words signify ideas or concepts originates from Aristotle, and that view was widely discussed by the late thirteenth century. (For Aristotelians, spoken words were signs of "concepts" in the mind.[26]) As other signification theorists, Descartes is not systematic about his use of the term "signification"; sometimes it is also the things represented by ideas that are signified instead of ideas.

In a theory of signification, the main (or maybe only) semantic relation words have is the *signification relation*. Words are considered as *signs* of ideas or conceptions in the minds of speakers, and occasionally also as signs of the public ordinary objects they are usually used to name or talk about.[27] As the nature of such signification is anything but clear, the theory has received much criticism. For example, J. S. Mill thought the theory amounts to holding that words *name* or refer to ideas (as opposed to ordinary things), which leads to some absurdities. In correcting what he perceived as mistakes of Thomas Hobbes, Mill writes: "When I say, 'the sun is the cause of day,' I do not mean that my idea of the sun causes or excites in me the idea of day."[28] Mill thinks that in Hobbes' use signification amounts to referring (as defined in the introduction above), so that, for Hobbes, words refer to ideas as opposed to ordinary objects. That Mill so thinks seems evident on the basis of his discussion. While arguing that there are good reasons for calling "the word *sun* the name of the

---

[24] CSM2 113; AT7 160.

[25] Descartes received a Jesuit education, and it was especially Jesuit philosophers who discussed and developed the theory of signification in their logic books; see Ashworth 1981.

[26] Ashworth 2012, 300.

[27] At least this is so for Descartes, Hobbes, Port-Royalians, and Locke.

[28] Mill 1974, 25; I.ii.1.

sun, and not the name of our idea of the sun," Mill also cites as evidence the fact that "names are not intended only to make the hearer conceive what we conceive, but also to inform him *what we believe*."[29] Though this formulation leaves enough room for debating, it seems natural to read Mill as pointing to a crucial difference between merely entertaining a conception in one's mind, on the one hand, and holding that conception to be true (or false), on the other. If this is what he means, then Mill must think that it is crucial that words name objects, not ideas, *because* it is objects, not ideas, that are relevant for the truth or falsity of our conceptions. And the name for such truth-relation between linguistic expressions and their truth-makers is *reference*.

The problem with Mill's criticism is that though Hobbes is less than clear how names signify our conceptions, he is quite clear that signification is not referring.[30] In *De Corpore* Hobbes first says that names "are signs of our conceptions" and "not signs of the things themselves."[31] But right after this he also states that some words like "a man, a tree, a stone," though not all of them, "are the names of the things themselves."[32] This is not the place to argue for an interpretation of Hobbes' views, but at least it seems clear that Hobbes was not guilty of the mistake Mill accused him of. Signifying and naming are distinct for Hobbes, though they sometimes can coincide. For Hobbes all names signify conceptions in the mind of the speaker, but at least some of them name ordinary things in addition.[33] One motivation for this view is the existence of empty names, that is, names that lack referents (an issue I discuss later in both sections 3 and 4). Empty names behave in the same ways in linguistic constructions as non-empty ones, and they can be used meaningfully despite their lacking referents in actuality. This is especially problematic for theories of direct reference, which seem to lack any plausible means to explain how empty names can be meaningful yet lack reference. As I mentioned in the introduction, my thesis

---

[29] Mill 1974, 24; I.ii.1; emphasis added.
[30] For an overview, see Duncan 2016.
[31] Hobbes, *De Corpore* 1839, I.ii.5.
[32] Hobbes, *De Corpore* 1839, I.ii.6.
[33] Hobbes, *De Corpore* 1839, I.ii.7.

that Descartes is a direct referentialist *and* also committed to the theory of signification might seem inconsistent because of this. No doubt many would find it more natural to think that the conceptions or ideas are signified precisely because they are like Fregean senses that mediate reference (in case there is an object to be referred to). I hope to show that the incompatibility between theories of signification and direct reference is only apparent, and not real.

I believe that Descartes would accept roughly the same view about signification relation as a mode of *signaling* (or indicating) that is articulated by Hobbes (and as analyzed later by Ott and Lowe).[34] Hobbes points out that "those things we call SIGNS are the antecedents of their consequents, and the consequents of their antecedents, as often as we observe them to go before or follow after in the same manner."[35] The example given by Hobbes elucidates the point nicely: "a thick cloud is a sign of rain to follow, and rain a sign that a cloud has gone before." Even if words are not natural but merely conventional signs, they are signs in this same sense. Words both signal the speaker's ideas for the hearer and the ideas signal which words the speaker must choose to convey her ideas. Signification—or linguistic signaling—is a matter of interplay between thoughts and public linguistic conventions. In this view, words are mere *tags* for ideas.

Once signification is understood in this way, the main thesis of the theory of signification becomes the following: "intentionality takes place at the level of ideas, not words."[36] Words are merely physical entities with nothing but physical properties, be they sounds, inscriptions, hand-signals, or flashes of light. They lack intrinsic meaning-related properties but can acquire conventional meanings by being associated with ideas. If we accept that—as Descartes and his followers at Port Royal did—"we can have no knowledge of what is outside us except by means of the ideas in us," then understanding language turns out to be mostly a matter of

---

[34] Ott 2003, chap. 1; Lowe 1995, chap. 7.

[35] Hobbes, *De Corpore* 1839, I.ii.2. See Ott 2003, chap. 1.

[36] Ott 2008, 294; see also 2003.

understanding how our thinking and its objects are connected.[37]

From the viewpoint of direct reference, however, this view of words being mere tags for *ideas* seems problematic, to say the least. Direct reference is often viewed as the view that names are tags for ordinary things, and that words lack other kinds of semantic content altogether.[38] This is how Mill, who is often considered as an early direct referentialist, seems to have viewed the opposition between the views. Precisely because names are like tags, they can be tags only for ordinary objects *or* their ideas, but not both. Combined with the view that our only access to extra-mental reality is by way of ideas, as the Cartesians have it, the tagging conception of names entails that tagging the ordinary objects is not a possibility—which is precisely the inconsistency of which Mill accuses Hobbes.

My argument in the coming sections is built around the attempt to show how Descartes' view of ideas and their features—objective contents and representationality—can escape the seeming inconsistency. In the remainder of this section, however, I want to say something about *subjectivism* concerning meaning.

One option that we can rule out in case of Descartes is his happily accepting subjectivism about meaning as a natural consequence of his views. There is plenty of evidence to the contrary. For instance, in replying to Hobbes's objections to *Meditations*, Descartes points out: "Who doubts that a Frenchman and a German can reason about the same things, despite the fact that the words that they think of are completely different?"[39] Regardless of any interpretational issues about signification, there is evidence that Descartes is at the least a firm believer in the publicity of meanings.

Of course, that Descartes is not committed to subjectivism as such does not mean his commitments would not entail it. Frege much later considered ideas as ill-suited to be bearers of public meanings precisely because he considered them

---

[37] Arnauld and Nicole 1996, 25.
[38] See, e.g., Marcus 1961, 310.
[39] CSM2 126, AT7 178–179.

necessarily subjective or "private."[40] This follows seemingly directly from his definition of "idea," which seems to be quite close to Descartes' view. Frege defines ideas as imperceptible by ordinary senses and as something "had" as contents of one's consciousness (i.e., sensations are not sensed themselves but had). More to the point, ideas depend on their subject and can belong only to one subject: "no two men have the same idea."[41]

Superficially, at least, Descartes seems committed to Frege's views about ideas. For Descartes, ideas are immaterial modes of the thinking substance (i.e., a mind), and modes are *states* or *ways* in which the substance exists at a moment.[42] Thus, two substances sharing the same mode is impossible.[43] From this substance-mode ontological viewpoint it seems that Descartes' theory of signification unavoidably leads to subjectivism about meaning, as Frege would argue. If ideas could be shared in the way public meanings must be, it would have to be possible for an idea to exist independently of a particular thinking subject. But because an idea is a state of a particular subject, dependent for its existence on that subject, ideas cannot be shared, and therefore, they cannot be what constitute or carry public meanings. Thus, Descartes' view that a German and a Frenchman or any other two speaker-thinkers could share meanings seems unwarranted by his own views.

I, however, think the above reasoning is flawed. I will next show how the conclusion that ideas cannot be shared does not follow from the view that ideas are states of a subject (and ontologically dependent on that subject).

---

[40] Frege 1956, 301–302.

[41] Frege 1956, 299–300.

[42] CSM1 201–212; AT8A 25–27.

[43] Though it has been argued (e.g., Hoffman 1990; see also Schmaltz 1992) that in certain cases (sensations and physical surfaces are cases in point) Descartes allows that two substances can share a mode. However, this issue has no bearing on the point I discuss in the text.

3. Objective reality

In certain writings at least, Bertrand Russell was no friend of subjectivism.[44] He suggested that propositions (related to the intersubjective meanings) can be considered in a way that I would like to offer as a model for understanding what ideas are for Descartes. In a famous letter to Frege, Russell writes the following:

> I believe that in spite of all its snowfields Mont Blanc itself is a component part of what is actually asserted in the proposition 'Mont Blanc is more than 4000 metres high'. We do not assert the thought, for this is a private psychological matter: we assert the object of the thought, and this is, to my mind, a certain complex (an objective proposition, one might say) in which Mont Blanc is itself a component part. If we do not admit this, then we get the conclusion that we know nothing at all about Mont Blanc.[45]

What Russell here labels "the thought" is what Frege called "an idea" (and what Russell calls "propositions" is what Frege calls "thoughts"). Both agree that such mental states are subjective, and thus cannot constitute meanings. However, Russell's point is that the threat of subjectivity is not the only problem we must worry about, for it cannot be the case Mont Blanc is irrelevant for the assertions naming it—the mountain itself must be involved in propositions concerning it *in propria persona*, so to say.[46]

If we construe propositions as distinct from their objects (as Russell thought Frege did), then there is nothing that could explain how those abstract meaning-entities are *about* the ordinary objects. This is the problem of intentionality—how do propositions come to have, or to be about, objects? Russell's point (one of many) here is that even if the problem of subjectivity of meaning is averted by postulating propositions as the intersubjective contents of thought, postulating them can involve a jump out of the frying pan into the fire. Without an intelligible connection to the propositions, the

---

[44] That is, in the relevant writings that are prior to his Russell 1918.

[45] *Russell to Frege 12.12.1904,* in Frege 1980, 169.

[46] The vague but expressive notion comes from Lovejoy (1923, 454) and is quoted by Hoffman (2002, 169).

ordinary objects named in the sentences expressing a proposition seem just irrelevant for the proposition. Therefore, to avoid both subjectivism of meaning and irrelevancy of propositions, Russell goes on to affirm a view that Frege found just as problematic as subjective meanings.[47] According it, propositions have parts that are not conceptual, but the physical worldly objects themselves that are named.[48] I will not discuss ontology of propositions further here, but I will later examine the competing conceptualist view held by Descartes that does away with such abstract objects altogether. For now, I wish to focus on how Descartes' view escapes both subjectivism about meaning and irrelevancy of objects by what he calls "*objective reality* of ideas" or "the *objective being*" (of the objects of ideas).[49]

According to Descartes, our ideas come to have objects, to *represent* or *intend* a thing outside the mind, by way of containing the *reality* (lat. *realitas*) or simply *the being* of that object.[50] Elsewhere, I defend an ontologically realist reading of the doctrine. There I argue that Descartes' claim, according to

---

[47] The example Russell takes up is in fact originally Frege's: "Truth is not a component part of a thought, just as Mont Blanc with its snowfields is not itself a component part of the thought that Mont Blanc is more than 4000 metres high" (Frege to Russell 13.11.1904, in Frege 1980, 163).

[48] In addition to the letter from 1904 quoted in the text, Russell expressed commitment to the view in his *Principles of Mathematics* (1903). But by the time of "On Denoting" (1905) he had already rejected it. Kaplan (2012) elaborates on the neo-Russellian framework of singular propositions. On the nature of Russellian propositions and direct reference, see also Wettstein 1986, 1990; and Almog 2012.

[49] The nature of Descartes' variety of objective being view is controversial. It is, for instance, the root of the debate about Descartes' commitment to direct or indirect realism, acting as the ground for totally opposite views. For important discussions on the direct realist side, see, e.g., O'Neil 1974; Yolton 1984; Normore 1986; Nadler 1989; Almog 2002, 2008a; Alanen 2003; and Brown 2007; on the indirect / representationalist side, see, e.g., Kenny 1968; Wilson 1978; Kaufman 2000; and esp. Hoffman 2002.

[50] Descartes' terminology makes it clear that the relation between an idea and the objective reality is one of containment (*continere*) or possession (*habere*). Objects are said to transfer or "'pour" (*transfundere*) their own reality into the ideas causally. (E.g., AT7 40–42; CSM2 28–29.) Ideas also exhibit the objective reality they contain, and this "objective mode of being belongs to ideas by their very nature" (CSM2 29; AT7 42).

which the objective reality is contained quite literally in the ideas yet identical with the objects, is contrary to some accusations, philosophically coherent.[51] We need not get into that discussion, for it suffices to note how Descartes' views about objective reality change depending on whether we are reading it as entailing direct reference or *direct perception*.[52] I argue here only that it entails the first (which is much less demanding a position than the latter, which seems to require that the object of thought is present in the mind by way of an idea *in propria persona*).

Descartes claims that the object itself is contained in the idea.[53] He writes:

> 'Objective being in the intellect' [- -] will signify the object's being in the intellect in the way in which its objects are normally there. By this I mean that the idea of the sun is the sun itself existing in the intellect—not of course formally existing, as it does in the heavens, but objectively existing, i.e. in the way in which objects normally are in the intellect.[54]

In my view, Descartes is trying to make sense of the idea that although ideas considered as modifications depend ontologically on the thinking substance they modify, they are nevertheless ontologically dependent also on the objects that cause the mind to be modified in that way. The connection between an idea of the sun and the (formally, i.e., actually existing) sun is an existential one, the former not being possible without the latter being related to the intellect in the right way. This two-way ontological dependence of ideas on both the subject as well as the object can be used to overcome the problem of subjectivism brought about by the fact that ideas are nevertheless modifications belonging only to one thinking

---

[51] Sinokki (forthcoming). For example, Yolton (1984) thinks that the objective containment is merely metaphorical and has no metaphysical import, and Kaufman (2000, 390) thinks the view makes 'no philosophical sense'. See also Hoffman 2002.

[52] See footnote 49 above.

[53] CSM2 75; AT7 102–103. *Pace* (e.g.,) Yolton 1984 and Kaufman 2000. For a careful analysis, see Hoffman 2002. I do not agree with Hoffman's conclusion that Descartes is an indirect realist, though.

[54] CSM2 75; AT7 102–103.

substance. Obviously, many minds can have modifications that are caused by the one and the same object.

This issue is very relevant for Descartes' view of perception, yet it has proven difficult to get the details straight. Though Descartes insists that it is the sun itself that exists in the intellect, the sun nevertheless has two distinct modes (ways) of being. Descartes thus seems to be committed to the view that in perception we are aware of the sun *only* in the objective sense, and that this objective sun *represents* the sun in its actual mode of being in the sky—which, if true, would be enough to view him as a representationalist.[55] Luckily for us, this problem about two modes of the sun doesn't really pertain to the current question about language and signification and we need not resolve it here. For our purposes it suffices to see that the connection between an idea and its object is a necessary one.

For Descartes, ideas come to have the objects they have because of their causal origination.[56] In fact, the only reason Descartes ventures into metaphysics of causation in the *Meditations* is to articulate how the objective reality of our ideas obeys the laws of ordinary causation. An idea of the sun is any idea that is caused by the sun. This amounts to the view that an idea of the sun involves *essentially* (necessarily) the sun itself; otherwise, it is not an idea of the sun at all but of something else.[57]

To elucidate, let's use the idea of the sun to consider the case of two distinct ideas, called I_1 and I_2, in two different scenarios. Let's stipulate that I_1 and I_2 are completely indistinguishable for the subject S (whose ideas are in question in both scenarios). In the scenario involving I_1, the idea originates in the sun in the way ideas ordinarily do. As a result of this origin, we can say that in this scenario it is the sun that objectively exists in S's mind when S entertains I_1. According to Descartes' view about necessity of causal origin of an idea, then, I_1 is an idea of the sun, and not of something else (i.e., I_1 is the sun itself existing in the intellect). Now, in the other

---

[55] This is essentially how Hoffman (2002) presents the case.

[56] CSM2 28, AT7 40–41.

[57] Kripke's (1980, 3rd Lecture) arguments for essentiality of causal origins thus apply to ideas as Descartes conceives them.

scenario, involving I_2, *S* is in the same position as the previous scenario. However, in this case I_2 originates not in the sun, but in the activity of an omnipotent deceiver (like the one introduced by Descartes in *the First Meditation*). So, in this scenario there is in fact no sky, no earth, and—it is worth emphasizing—no sun, *nor has there ever been*. There are only demon-caused hallucinatory experiences in the mind of *S*. In this latter case, when I_2 is in *S*'s mind, that which exists objectively in *S*'s mind has nothing to do whatsoever with the sun. Therefore, in line with Descartes' view about necessity of causal origin of an idea, I_2 is not an idea of the sun at all, despite being indistinguishable from one.

What I take to be the important point in Descartes' theory of objective being or reality is this: In the above example, at best, I_2 is a *fake* idea of the sun. Just like a fake gun cannot be used to shoot bullets, a fake idea cannot be used to think of the sun. The connection between an idea and its object is essential (necessary) for the idea in question precisely because in order to think of the sun, you need an idea that objectively contains the sun. An idea not containing the sun does not allow thinking of the sun, but only something else that, at best, has the appearance of the sun.

As for Frege's concern about ideas leading to subjectivism, Descartes' view seems to defuse it quite thoroughly. Though it is impossible for two subjects to share an idea in the sense of sharing a modification belonging to a particular mind, two minds can nevertheless be modified by the same object. This amounts to two subjects having the same idea in their minds (*pace* Frege), and in a sense that is metaphysically just as important as the substance-mode ontological sense—both the subject whose modification is in question as well as the object that is the causal origin of that modification are just as essential for the idea.

Importantly to our discussion, because names tag ideas, and ideas necessarily objectively contain their originating objects, there really cannot be any alteration in references of names either: when a name signifies an idea containing objective reality of an object, O_1, it thereby refers to O_1. This seems to take care of Russell's worry about irrelevance of ordinary objects. What is more, standard externalist considerations

presented by Kripke and Putnam seem to apply here.[58] A person speaking English refers to sun every time they utter the expression "the sun," for the conventions of that linguistic community dictate that the expression always signifies an idea that contains objectively the sun. Of course, it is possible to signify idiosyncratically some other ideas by the expression "the sun" but such signification amounts to making a linguistic mistake—it is a case of using a name that does not name the object to which one tries to apply the name. Furthermore, that kind of Humpty Dumpty use of words does not amount to genuine language use in the sense Descartes understands it, for one can communicate one's ideas successfully only if the linguistic conventions of the public language in question are observed sufficiently. Finally, in the case imagined above where *S*'s perceptions consist of hallucinatory experiences produced by an omnipotent deceiver, *S* would not even speak English, for none of her ideas contain objectively the things words of English signify.[59]

When his view about objective reality is understood as I have presented it, Descartes' view of signification combined to his theory of ideas amounts to a theory of direct reference (about ordinary, singular objects, that is). As for what kind of direct reference this view precisely amount to, that can be answered only after examining the other aspect of ideas—their representational character.

Before moving on, however, I would like to address an objection that my interpretation might elicit.[60] If it is true that all our thoughts and knowledge of things proceeds by way of ideas, and names are mere tags for those ideas, then it is plausible to ask how do we know that "the sun" is a tag of the same idea for you and me? If the only answer we can provide is (as my appeal to causal origination of ideas seems to imply) that we know it in the same way as we know that the expression refers to the sun, then the ideas seem to do no work in explaining the workings of language.

---

[58] Kripke 1980; Putnam 1975. See also Raatikainen 2020 and Haukioja 2017.

[59] This demon-case seems in many ways analogous to Putnam's (1975) Twin Earth case.

[60] I am grateful for Tapio Korte for drawing my attention to this issue.

There are several related points that can be used to counter the objection. First, notice the problem of empty names. Names like "Vulcan" that lack referents are problematic for direct reference. Those names behave linguistically just like ordinary names that have referents and can be used to convey meaningful thoughts, yet they do not refer to anything. One benefit of seeing ideas as an ingredient in semantics lies in explaining the behavior of such empty names—this is in fact one of the main reasons also Hobbes cites for thinking that all names signify conceptions in the minds.[61] In the reconstructed signification theory I attribute here to Descartes, empty and non-empty names do not differ linguistically. What precisely is empty is the idea, not the name—there can be no such objective reality as the reality of planet Vulcan, for such a planet does not exist and cannot cause any ideas in us. However, there still is an idea signified by the name "Vulcan," but it is a fake idea of planet Vulcan in the sense discussed above. It appears like an idea of a planet, but it cannot be used to think about an actual planet.

In my view, though I will not argue further for this here, Descartes' famous example of the intricate machine shows that Descartes sees invented ideas as having composite objective realities.[62] Invented ideas do not contain the objective reality of any one particular thing, for their objective reality is a patchwork of pieces from diverse sources. Such ideas nevertheless have ordinary representational characters (see shortly below), which explains why cognitively those ideas can also appear like ordinary ideas (e.g., compare the astronomers' idea of the sun Descartes discusses, quoted above, to the empty idea of Vulcan; both are products of similar astronomical reasonings).

Another aspect of why words must signify ideas is related to communication. As stated above, for Descartes, language is a system which enables speakers to encode their thoughts into physical representations (noises, patterns, sign marks…) that can be decoded at the receiving end by the audience. The exchange of such physical signs is characterized by Descartes occasionally as the "passing of an idea" from one thinker to

---

[61] Hobbes, *De Corpore* 1839a, I.ii.6.

[62] CSM2 75, AT7 104. See also Sinokki 2016, ch.3.

another.[63] As arguably even a group of parrots (for details, see reference in the footnote) could pass on such physical marks from parrot to parrot while still not passing on any semantic information or meanings (ideas) whatsoever, we need to add into the picture something carrying the meanings that is transmitted in cases of genuine language-use.[64] We saw Russell claiming that what carries such meaning is a proposition; for Descartes it is an idea containing an objective reality. It could be even argued that at least *prima facie* sharing of thoughts or ideas by transmitting physical marks (produced by our tongues and received by our ears) is significantly *less* problematic than the claim that in addition to this, certain things called propositions (that cannot be touched or be seen) are involved in the business.

Be that as it may, propositions were important for Russell among other things because of their *structuredness*.[65] In contrast to their individual constituent parts (e.g., concepts like "white" and objects like Mont Blanc), he considered propositions as structured unities that bear meanings. A proposition is, in this sense, something more than a mere collection or list of things. It is a "complex" that (conceptually or logically) organizes things into relations and represents things (or states of affairs) as being in this or that way. This unity and logical structure are what make the analysis of such things possible. Next, I will argue that ideas considered from the cognitive aspect of ideas that I call *representational character* can perform this conceptual role Russell (and Frege) thought requires postulating propositions.

4. Representational character

As Margaret Wilson expresses in frustration, Descartes' view of ideas "entails that the objective reality of an idea is not *something the idea wears on its face*."[66] As we saw, ideas I_1 and I_2

[63] CSM2 11, AT7 14–15.

[64] For a sustained elaboration of this thought-experiment in context of Kripke's causal theory of reference, see Sinokki 2022.

[65] A caveat must be stated; when discussing Russell, I mean to make statements only about contents of the specific works already cited, so I do not intend to generalize.

[66] Wilson 1978, 98.

can contain totally unrelated objective realities yet still be subjectively indistinguishable to the thinker. To understand how this is possible we must take into account that Descartes characterizes ideas as thoughts that are "as it were the images of things."[67] To capture what it properly is that ideas wear on their face, like images, Wilson coins the notion of *representational character* of an idea.[68] Wilson ultimately finds this divorce between the objective reality of an idea and its representational character "an embarrassment, not an asset."[69] I strongly disagree with this assessment, for, on the contrary, I see this divorce as the major strength of Descartes' view. I believe (though I won't argue for it here) that the reason for Wilson's disappointment is that she, like many other commentators, gets the relation between the two backwards. In her view, the representational character must determine the object of the idea, and once she sees, quite correctly, that for Descartes it is instead the objective reality that determines the object of the idea, she finds the view incoherent.

As we noted at the end of last section, for Russell a proposition was a structured unity that presents things or states of affairs as being in this or that way; importantly, proposition is not a mere collection or a list of things but a precisely a *structured unity*. Just like images (ignoring abstract art for the moment), propositions also present a single view of what they present. Moreover, images and propositions do this in virtue of the arrangements of their constituent parts. How their parts are related to each other matters for how things are represented as being. In my view, the kind of representational character we can attribute to Descartes amounts to the way in which the objective contents are arranged in, or presented by, the idea. In my view, it is precisely in this structural, conceptual sense that ideas are *as if images* for Descartes.[70]

In *the Second replies*, Descartes defines ideas as "the form of any given thought, immediate perception of which makes me

---

[67] CSM2 25, AT7 37.

[68] Wilson 1978, 90.

[69] Wilson 1978, 98.

[70] Cf. Wilson 1978, 89ff., who discusses representational character especially in relation to sensations and connects it to phenomenality rather that concepts.

aware of the thought."[71] As Cottingham points out, in part because of their formal features "Cartesian ideas are in some respects much more like publicly accessible concepts than private psychological items."[72] The form of thought is the "structure" of the idea which presents things as falling under concepts. These forms are something that can be instantiated in several minds and in several ideas. It is this form or the representational character of the ideas that is indistinguishable in ideas I_1 and I_2 above. Due to space limitations, I will here restrict my attention to the representational character as something conceptual and ignore altogether "qualitative" aspects of it (e.g., phenomenality) for the irrelevance of the latter for the purposes of this paper.

In my view, the representational character is a *mode of presentation* of the objective reality contained in the idea. However, it is not at all like a Fregean "sense" in being a mode of presentation which determines an object. As we saw in the previous section, the idea as a modification of a thinking substance is also a product of the object. That the objective reality comes to mind is a matter of causation. Now, that this objective reality is presented in this or that way similarly flows from the causal connection to the object and does so in accordance with the vagaries of the relation we happen bear to the object. (The way distance affects the visual and auditory appearances of things is an example of such vagaries.)

All this talk of conceptual structure of ideas makes more sense when we consider the fact that Descartes is a *conceptualist* about universals and abstracta, such as mathematical objects.[73] That is, for Descartes there is nothing general or universal outside *any* mind, but plenty that is so within all

---

[71] CSM2 113; AT7 160. According to Descartes, "thought" is used to refer to "everything that is within us in such a way that we are immediately aware of it" (CSM2 113; AT7 160).

[72] Cottingham 1997, 39. For Cottingham, these formal features of thought are naturally connected to what Descartes in the Fifth Meditation calls "true and immutable natures," which I will not discuss here due to space constraints.

[73] With ample textual evidence, there is a good case to be made in favor of attributing thoroughgoing conceptualism to Descartes. This view has been elaborated and defended most notably by Lawrence Nolan in a series of papers; see Nolan 1997a; 1997b; 1998; 2011; 2015; 2017.

minds. According to Descartes, all "eternal truths" reside "within our mind" and "[n]umber and all universals are simply modes of thinking."[74] Such universal ideas are formed by cognitive processing, by abstraction and exclusion, for instance.[75] In general, Descartes thinks that these ideas are innate to the mind in the sense of not requiring extra-mental causes like singular ideas, discussed above, require.

This gives us a clue as to how objective reality and representational character relate to one another. While objective reality is something coming into the mind from the outside causally (and can be informationally rich or meager depending on the vagaries of the occasion), universals are conceptual forms by which the mind reacts to that incoming thing with the result that the thing is presented to the mind as being in this or that way. For instance, the sun can come to exist in my mind through its causal action on my senses. It is thereby presented as round and light-emitting, properties which appear as forms which I can abstract from that idea. Evidently for Descartes, such representational characters of ideas are often not quite static but can change in response to our reasoning processes and if they are considered in conjunction with other ideas. Without entering this complex topic, those generic or abstract representational characters that, in contrast, do not change at all Descartes calls "*true and immutable natures*."[76]

---

[74] CSM1 208–209, AT8A 22–23; CSM1 212; AT8A 27. Clearly, Frege's insistence that a "third realm must be recognized" (1956, 302) cuts no ice inside an ontology like this, for it mustn't.

[75] CSMK 236, AT4 120; Murdoch 1993; Nolan 1997a. Descartes conceives abstraction in terms of selective attention to a particular aspect of an idea, while exclusion is the active denial of an aspect of an idea; see Nolan 1997a, 133.

[76] In my view true and immutable natures are conceptual entities, existing only in the mind, as Nolan (1997b) argues. Along with Nolan, I believe that true and immutable natures are realities that can exist only in the mind, from which it follows that for abstract ideas, they are also the objective realities contained in those ideas. It is important to notice that for Descartes the fact that true and immutable natures "do not depend" on one's mind does not mean that their existence would not depend on the existence of thought more generally. In my view invented ideas, like ideas of chimera, have composite objective realities, gotten from diverse

In *the Third Meditation*, Descartes asks us to consider two ideas he has of the sun. One idea originates in the sense-perception of the sun, while the other is based on astronomical reasoning. The visual idea, he writes, "makes the sun appear very small," while the reasoning-based, intellectually constructed idea "shows the sun to be several times larger than the earth."[77] Descartes points out: "Obviously both these ideas cannot resemble the sun which exists outside me; and reason persuades me that the idea which seems to have emanated most directly from the sun itself has in fact no resemblance to it at all."[78] As representations, the ideas are very different. That they are ideas of the same thing (the sun) is determined by the fact that they both contain the *objective reality* of the sun. But their difference makes it very clear that the objective reality of an idea is not something the idea "wears on its face"; the objective reality is simply that which can be represented in different ways. And finally, though representational character is obviously a mode of presentation here for the sun, it cannot be what determines the object of the idea. If it were, the visual idea that emanated most directly from the sun would not have the sun as its object, but at best some much smaller yellow disc (which is precisely what Descartes denies being the case here).

Now, how does representational character fit together with signification? Consider first the case of names of mathematical objects. As according to Descartes, such things exist only in the mind—not as modifications of a particular mind, but as features of thought in general—the relations of signifying and reference will coincide in this case just as they did in the case of ideas containing things objectively (see the previous section). Descartes' example of an idea of a *chiliagon* offers a nice illustration. According to Descartes, a mentally visualized image representing a chiliagon is confused and obscure, and it cannot be distinguished from mental images of other similar figures with very many sides. Still, our understanding of

---

sources, while nevertheless conceptually the chimera (i.e., the representational character of an idea of chimera) can have a true and immutable nature.

[77] CSM2 27; AT7 39.

[78] CSM2 27; AT7 39.

the chiliagon is clear and distinct, for we can demonstrate mathematically many things of the figure.[79] Here we can say that though the representational character of the idea of a chiliagon constructed in the imagination is confused, the chiliagon that is contained objectively in that idea, and is grasped by the understanding, is what is properly signified by the name.[80] In case of ideas of universals that apply to several things, Descartes says that "we apply one and the same term to all the things which are represented by the idea in question, and this is the universal term."[81] In both cases the object referred to (a mathematical object or a universal) exists only as a form of thought. That object is what is contained in the idea signified by the name as well. Therefore, even here signification and reference coincide.[82]

Invented ideas (discussed already briefly in connection to objective reality above) are akin to ideas of mathematical objects and universals. However, it is important to notice, as Nolan has argued, that the distinction between the two is crucial for Descartes.[83] Invented ideas, like those of chimeras, originate in the mental activity of the thinker who combines ideas into new complex arrangements.[84] This is why those ideas lack a singular objective reality and are patchworks of

---

[79] CSM2 50, AT7 72; CSM2 264, AT7 384–385.

[80] Nolan (1997b) argues that the universals, having existence only in thought, are also thus the objective realities contained by those ideas. Though I agree with Nolan's argument in principle, I somewhat hesitate to accept the conclusion. Objective reality is for Descartes clearly something obeying ordinary causation, and I am not sure that the formal–conceptual entities, such as universals must be, are apt to obey causation in the required sense. This problem must be addressed properly on another occasion.

[81] CSM1 212; AT8A 27. This issue relates to Descartes' conceptualism, as discussed below.

[82] Notice that the ontology required by this view is not, *prima facie*, any more problematic than the seemingly Platonic abstract entities somewhere outside the mind, to the existence of which Frege and Russell are committed.

[83] Nolan 1997b.

[84] Of course, that idea is not created *de novo* every time someone thinks of it, but rather "passed on" from the inventor onwards. The similarities of Kripke's causal transmission of names and Descartes' causal "passing on" of ideas are evident, and a topic for another paper.

several distinct realities. Ideas of mathematical objects and universals are not like this, for their objects have—or are—forms that are independent of any individual thinkers (despite existing only in thought). Yet both kinds of ideas have representational characters. As I pointed out in the last section in relation to empty names, ideas of chimeras and the like lack referents, for they do not contain singular objective realities. Insofar as we can consider mathematical objects and universals as singular beings (though existing only in thought), we can say they are the referents of the names that signify the corresponding ideas. In this, names of mathematical objects and universals are more like names of ordinary singular objects, and unlike empty names lacking such singular actual referents. However, discussing this complicated issue further is not possible in this connection.

In my view, it is precisely the interplay between objective reality and representational character that solves many traditional puzzles that create problems for direct reference theories. In Descartes' view, a thinker might have two ideas with the same objective reality, but with so different representational characters that she is not able to realize that those ideas are but two different representations of one and the same thing. Seeing Venus in the morning sky and then again in the evening sky would be a case in which, due to the vagaries of the situation, an (ancient) astronomer could have had two ideas of one single object without realizing that there is only one thing (just as he didn't realize that what he sees is not a star but a planet). Signifying those ideas with different names like "the Morning star" and "the Evening star'" could eventually result in a significant discovery of the fact that what we thought of as two distinct stars was in fact only one. But for Descartes, this discovery is not about the names any more than finding out that the thing is not a star but a planet. It's a realization about our conceptions or ideas, and how what we know relates to things our ideas are about. As regards reference, it has all along been direct. Despite several names, only one thing, Venus, has been involved all the time. The distinct names were tags for ideas containing one and the same objective reality all along. Yet due to the vagaries of the situation, *qua* representations of the second rock from the sun, the ideas were so confused and obscure that the realization that they

were ideas of the same object required highly sophisticated astronomical reasoning.

One further point about representational character and how it is determined by what I have vaguely referred to as "vagaries" of the situation in which an objective reality is gotten into the mind. That an idea always has a representational character also allows for cases of radical misidentification, like the cases envisaged by Keith Donnellan.[85] Donnellan presents a case in which *S* thinks she sees the history professor, but in fact the thing *S* sees is just a rock in the shadows. In this case, the idea is of a rock. Yet due to the perceptual situation, the representational character of the idea triggers a judgment that it is the history professor there. Consider now *S*'s following soliloquy. Seeing something in the shadows, *S* utters: "What is *that*?" After an inconclusive peer into the darkness, *S* replies to herself: "That's got to be *the history professor*!" It is easy to see that the italicized expressions in these quotes do not refer to the same thing nor do they signify the same idea. In both sentences, "that" refers to the stone and signifies the confused perceptual idea of it. In the latter sentence, "the history professor" refers to the history professor, and signifies an idea of that person, who is mistakenly identified with the stone.[86] Though I cannot go into the details fur-

---

[85] Donnellan 1966, 295ff.

[86] As I see it, according to Descartes' view, the mistaken judgment expressed by sentence "That's got to be *the history professor*!" is not an identity judgment (i.e., a judgment of the form "*a* = the *F*"), but rather an attempt to predicate the property of "being the history professor" of the subject that happens to be the stone. Similarly for the question Russell attributes to George IV: It is queried of Scott, by signifying with his name an idea containing his objective reality, whether the property of "being the author of *Waverley*" can be truly attributed to him. Here "the author of *Waverley*" is a description connected to the representational character of an idea, a mode of presentation for a person. As only one person at best can be *the* author of *Waverley*, that representational character can truly go together only with ideas that objectively contain the person who actually wrote *Waverley*. The truth of the judgment or statement then depends on whether the objective realities contained in the ideas with different representational characters (i.e., of a person whose name is "Scott" and of the person who wrote *Waverley*) are the same or not; or in other words, whether the one idea can be truly affirmed of the other or not.

ther, I believe investigating this intriguing interplay between objective reality and representational character can be of help in understanding how language, thinking, and the world beyond these two properly all interlock together, as they obviously do.

5. Conclusion

I have argued that Descartes' views of ideas and signification together entail a picture of language that is directly referential. Words *signify* ideas, and though reference and signification must not be conflated, in many cases they coincide for Descartes. But seeing that this is so depends on a proper understanding of his metaphysics. Important is the causal relation between an object and its idea, requiring an essential (necessary) connection between an idea and the object causally originating it, and the conceptualist ontology that explicates the contents for ideas of universals and abstracta. Once the metaphysics is understood properly, the representational characters of ideas turn out to be mostly conceptual in their nature, and to function as modes of presentations of objects. Because the ideas have both objective and conceptual contents that are very much intersubjective—they are ordinary objects and conceptual universal forms, respectively—the meanings of language in Descartes' view are hardly at the risk of turning out to be "private."

The reason why the ideas are needed in explaining the workings of language *despite* direct referentiality of names is that just as ideas do not wear their objective realities on their faces, names do not wear their referents on their sleeves. It is the representational character of ideas, signified by words, that explains why empty names appear just like referring ones, or why some identity statements are not at all trivial, and so on. This makes a plenty of room for mistakes even in case of ideas that do contain actual singular objects objectively.

In closing, I would like to emphasize how direct reference is, in general, completely silent about the possibility of there being other kinds of semantic or informational contents besides the reference determining modes of presentation which

it forbids.[87] For instance, in astronomy variable stars (i.e., stars whose brightness varies) are named in accordance with a variation of Bayer designation format, a convention of naming that reveals the constellation to which the star belongs and the rank of the star in the order of their discovery. For instance, "UY Scuti" names the 38th variable star discovered in the constellation of Scutum. Thus, besides referring to the extreme red hyper-giant once considered being the biggest star discovered, the name also carries other kinds of highly meaningful information as well. Such information hardly amounts to a Fregean sense in determining the reference, for the name follows from vagaries of the astronomical research and the convention about naming variable stars. Before 1860 when Astronomers at Bonn discovered the star, the name did not determine this specific star, so it could have named which ever star would have been found next.[88] Still, the additional information carried by the name deserves the name *meaning* (it's even expressible as a definite description that is coreferential with the name!). So, direct reference must be considered a thesis about reference only, not a denial of other

---

[87] Recanati (1993) defends this claim in length. Of course, many proper names are also common names and can connote, say, biblical figures, and so on. But some authors, like Kaplan, would treat all such cases strictly as mere homoforms, as semantically distinct words which happen to have the same spelling, so such examples would not serve my goal. The example of Bayer designation format, however, is suited to my purposes, because it makes the additional information an essential part of the relevant naming convention (which, then, also affects the modal properties of the names).

[88] I think this example reveals nicely the mistake some, like Wiggins (2001, 132) and Noonan (2014, 144), make in criticizing Kripke's claim about the necessity of origin (1980, 112–13). They claim that the necessity of origin is shown problematic by examples about coreferential names and descriptions, such as Wiggins' claim that while intuitively Julius Caesar might not have a different father, quite intuitively the man whom Brutus murdered in 44 BC could have had a different farther. However, just as the name "UY Scuti" could have named another star, the description "the man whom Brutus murdered in 44 BC" could have picked up a man distinct from the one it *actually* picks up. That is, the objection is not about modal properties of things (contrary to Kripke's original point), but only about modal properties of the expressions used to designate those things. Therefore, the objection misses the point.

kinds of semantic relations or contents beyond reference. Thus, the representational character of ideas, too, can be considered as additional meaning-contents insofar as we are clear that this content has nothing to do with how the reference of name originally was determined.[89]

To my mind, the combination of direct referentiality and modes of presentations that do nothing to fix the reference but are highly useful in many other ways is not yet appreciated enough among philosophers of language and mind. Consider how direct referentialist David Kaplan, for example, is known for his thesis "No mentation without representation!"[90] Though Kaplan himself is ambivalent whether this means that a representation determines (always, sometimes, ever) the object of 'mentation', his critics often base their criticisms on the assumption that he thinks it so determines.[91] The view I am attributing to Descartes offers an interesting way of concurring with Kaplan's thesis without falling prey to its criticisms: it shows how a direct referentialist can be robustly realist about representational mental contents without there-

---

[89] I have not discussed here the possibility of fixing the reference of a name by using a description. I, however, have argued elsewhere that such fixing is in fact impossible (see Sinokki 2022).

[90] Kaplan 2012, 153. See also Almog 2005; Eaker 2004.

[91] For example, Eaker 2004, 381; Almog 2005, 520; 2014, 45; Stalnaker 2009, 233. In opposition, Bianchi (2007) points out that for Kaplan, representation can be taken as a vehicle of cognition, which does not determine the object.

Especially in his later works, Kaplan's remarks reveal that his view, in fact, is closer to Descartes' view than to that which Kaplan's critics attribute to him. True, Kaplan thinks that a "representation determines the referent," but only in the sense that it "leads to" (Kaplan's term) the referent; not in virtue of satisfaction conditions, but "by way of its origin, by way of a particular descending path through a network of *tellings about*, a path that ideally is ultimately grounded in an event involving a more fundamental epistemological relation" (Kaplan 2012, 153; see also 167, endnote 22). This seems to amount to similar causal connection I see as obtaining between the extramental object, its idea, and the word used to name the object (contained objectively by the idea that is signified). "Determination" in the sense Kaplan seems to have in mind, is not a satisfaction relation, but a two-way 'pointing' relation, much like the signification relation (see section 2 above).

by yielding neither to Fregean descriptivism nor the Platonism often associated with such Fregean view.

University of Oulu

References

Alanen, Lilli (2003). *Descartes's Concept of Mind*. Cambridge, MA & London: Harvard University Press.

Alanen, Lilli (2008). "Descartes' Mind-Body Composites, Psychology and Naturalism." *Inquiry* 51 (5), 464–84.

Almog, Joseph (1985). "Form and Content" 19 (4): 603–16.

Almog, Joseph (2002). *What Am I?* Oxford: Oxford University Press.

Almog, Joseph (2005). "Is a Unified Description of Language-and-Thought Possible?" *Journal of Philosophy* 102, 493–531.

Almog, Joseph (2008a). *Cogito? Descartes and Thinking the World*. Oxford: Oxford University Press.

Almog, Joseph (2008b). "Frege Puzzles?" *Journal of Philosophical Logic* 37(6), 549–74.

Almog, Joseph (2012). "Referential Uses and the Foundations of Direct Reference." In Joseph Almog and Paolo Leonardi (eds.), *Having in Mind*. Oxford and New York: Oxford University Press, 176–84.

Almog, Joseph (2014). *Referential Mechanics: Direct Reference and the Foundations of Semantics*. Oxford University Press.

Arnauld, Antoine, and Pierre Nicole (1996). *Logic or the Art of Thinking*. (Edited by Jill Vance Buroker.) Cambridge: Cambridge University Press.

Ashworth, E. J. (1981). "'Do Words Signify Ideas or Things?' The Scholastic Sources of Locke's Theory of Language." *Journal of the History of Philosophy* 19(3), 299–326.

Ayers, Michael (1998). "Ideas and Objective Being." In edited by Daniel Garber and Michael Ayers (eds.), *The Cambridge History of Seventeenth-Century Philosophy*. Cambridge: Cambridge University Press, 1062–1107.

Bianchi, Andrea (2007). "Speaking and Thinking (Or: A More Kaplanian Way to a Unified Account of Language and Thought)." In M. Beaney, C. Penco, and M. Vignolo (eds.), *Explaining the Mental: Naturalistic and Non-Naturalistic Approaches to Mental Acts and Processes*. Newcastle: Cambridge Scholar Publishing, 13–32.

Brown, Deborah (2007). "Objective Being in Descartes: That Which We Know or That By Which We Know?" In Henrik Lagerlund (ed.), *Repre-*

sentation and Objects of Thought in Medieval Philosophy. Aldershot: Ashgate, 133–51.

Burge, Tyler. (2012). "Referring De Re." In *Having in Mind*, edited by Joseph Almog and Paolo Leonardi, 107–21. Oxford University Press.

Capuano, Antonio. (2015). "Thinking about an Individual." In *On Reference*, edited by Andrea Bianchi, 147–72. Oxford: Oxford Univesity Press.

Chomsky, Noam. (1991). "Linguistics and Descartes." In *Historical Foundations of Cognitive Science*, edited by J-C. Smith, Philosophi, 71–79. Springer.

Clemenson, David. (2007). *Descartes' Theory of Ideas*. London: Continuum.

Cottingham, John. (1978). "'A Brute to the Brutes?': Descartes' Treatment of Animals." *Philosophy* 53 (206): 551–59.

—. 1997. "'The Only Sure Sign …': Thought and Language in Descartes." *Royal Institute of Philosophy Supplement* 42: 29–50.

Descartes, René (1904). *Oeuvres de Descartes*. (Edited by Charles Adam and Paul Tannery.) Paris: Léopold Cerf.

Descartes, René (1984). *The Philosophical Writings of Descartes: Volume 2*. (Edited by John Cottingham, Robert Stoothoff, and Dugal Murdoch.) Cambridge: Cambridge University Press.

Descartes, René. (1985). *The Philosophical Writings of Descartes: Volume 1*. (Edited by John Cottingham, Robert Stoothoff, and Dugal Murdoch.) Cambridge: Cambridge University Press.

Descartes, René (1991). *The Philosophical Writings of Descartes: Volume 3*. (Edited by John Cottingham, Robert Stoothoff, Dugal Murdoch, and Anthony Kenny.) Cambridge: Cambridge University Press.

Donnellan, Keith S. (1966). "Reference and Definite Descriptions." *The Philosophical Review* 75 (3): 304.

Dummett, Michael (1973). *Frege: Philosophy of Language*. London: Duckworth.

Duncan, Stewart (2016). "Hobbes on Language: Propositions, Truth, and Absurdity." In *The Oxford Handbook of Hobbes*, edited by A. P. Martinich and Kinch Hoekstra. New York: Oxford University Press.

Eaker, Erin L. (2004). "David Kaplan on De Re Belief." *Midwest Studies In Philosophy* 28 (1): 379–95.

Frege, Gottlob (1956). "The Thought: A Logical Inquiry." *Mind* 65 (259): 289–311.

Frege, Gottlob (1980). *Philosophical and Mathematical Correspondence*. (Edited by Gottfried Gabriel, Hans Hermes, Friedrich Kambartel, Christian Thiel, Albert Veraart, Brian McGuinness, and Hans Kaal.) Oxford: Basil Blackwell.

Harrison, Peter (1992). "Descartes on Animals." *The Philosophical Quarterly* 42 (167), 219–27.

Haukioja, Jussi (2017). "Internalism and Externalism." In Bob Hale, Crispin Wright, and Alexander Miller (eds.), *A Companion to the Philosophy of Language*. 2nd ed. Chichester: Wiley Blackwell, 865–80.

Hobbes, Thomas (1839). "Elements of Philosophy, First Part." In *The English Works of Thomas Hobbes, Vol. I.* (Edited by William Molesworth.) London: John Bohn.

Hoffman, Paul. (1990). "Cartesian Passions and Cartesian Dualism." *Pacific Philosophical Quarterly* 71(4), 310–33.

Hoffman, Paul (2002). "Direct Realism, Intentionality, and the Objective Being of Ideas." *Pacific Philosophical Quarterly* 83(2), 163–79.

Kaplan, David (1989). "Demonstratives." In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press., 481–563.

—. (2012). "An Idea of Donnellan." In Joseph Almog and Paolo Leonardi (eds.), *Having in Mind*. Oxford and New York: Oxford University Press, 122–75.

—. (2013). "De Re Belief." In Richard T. Hull (ed.), *Presidential Addresses of The American Philosophical Association 1981–1990*. Dordrecht: Kluwer Academic Publishers, 25–37.

Kaufman, Dan (2000). "Descartes on the Objective Reality of Materially False Ideas." *Pacific Philosophical Quarterly* 81 (4), 385–408.

Kenny, Anthony (1968). *Descartes: A Study of His Philosophy*. New York: Random House.

King, Peter (2007). "Rethinking Representation in the Middle Ages." In *Representation and Objects of Thought in Medieval Philosophy*, edited by Henrik Lagerlund, 81–100. Aldershot and Burlington: Ashgate.

Korte, Tapio. (2022). "Frege's Answer to Kripke." *Theoria* 88, 464–79.

Kripke, Saul. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lovejoy, A. O. (1923). "'Representative Ideas' in Malebranche and Arnauld." *Mind* 32(128), 449–61.

Lowe, E. J. (1995). *Locke on Human Understanding*. London: Routledge.

Marcus, Ruth Barcan (1961). "Modalities and Intensional Languages." *Synthese* 13(4), 303–22.

Mill, John Stuart. (1974). *A System of Logic, Ratiocinative and Inductive. Collected Works of John Stuart Mill, Vols. 7–8.* (Edited by J. M. Robinson.) London: Routledge & Kegan Paul.

Murdoch, Dugald Jan (1993). "Exclusion and Abstraction in Descartes's Metaphysics." *Philosophical Quarterly* 43(170), 38–57.

Nadler, Steven. (1989). *Arnauld and the Cartesian Philosophy of Ideas*. New Jersey: Princeton University Press.

Nolan, Lawrence (1997a). "Reductionism and Nominalism in Descartes' Theory of Attributes." *Topoi* 16(2), 129–40.

—. (1997b). "The Ontological Status of Cartesian Natures." *Pacific Philosophical Quarterly* 78, 169–94.

Noonan, Harold (2014). *Routledge Philosophy GuideBook to Kripke and Naming and Necessity.* Abingdon: Routledge.

Normore, Calvin G. (1986). "Meaning and Objective Being: Descartes and His Sources." In Amélie Oksenberg Rorty (ed.), *Essays on Descartes' Meditations*. Berkeley: University of California Press.

Nuchelmans, Gabriel (1983). *Judgement and Proposition: From Descartes to Kant*. Amsterdam: North-Holland.

O'Neil, Brian E. (1974). *Epistemological Direct Realism in Descartes' Philosophy*. Albuquerque: University of New Mexico Press.

Ott, Walter (2003). *Locke's Philosophy of Language*. Cambridge: Cambridge University Press.

Ott, Walter (2008). "Locke on Language." *Philosophy Compass* 3(2), 291–300.

Putnam, Hilary (1975). "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7: 131–93.

Raatikainen, Panu (2020). "Theories of Reference: What Was the Question?" In Andrea Bianchi (ed.), *Language and Reality from a Naturalistic Perspective: Themes from Michael Devitt*. Cham: Springer, 69–103.

Read, Stephen (1977). "The Objective Being of Ockham's Ficta." *Philosophical Quarterly* 27: 14–31.

Recanati, François (1993). *Direct Reference. From Language to Thought*. Oxford: Blackwell.

Russell, Bertrand (1903). *The Principles of Mathematics*. Cambridge: Cambridge University Press.

Russell, Bertrand (1905). "On Denoting." *Mind* 14(56), 479–93.

Russell, Bertrand (1918). "The Philosophy of Logical Atomism." *The Monist* 28(1), 495–527.

Schmaltz, Tad M (1992). "Descartes and Malebranche on Mind and Mind-Body Union." *The Philosophical Review* 101(2), 325.

Sinokki, Jani (2011). *Kaksi Kielifilosofia: Locke Ja Kripke.* ("Two philosophers of language: Locke and Kripke.") Masters Thesis: Turun yliopisto.

Sinokki, Jani (2016). *Descartes' Metaphysics of Thinking*. Turku: Turun yliopisto.

Sinokki, Jani (2022). "'What on Earth Is Smenkhkare?' WH-Questions, Truth-Makers, and Causal-Informational Account of Reference." *Theoria* 88(2), 326–47.

Sinokki, Jani (forthcoming). "Having a Cake and Eating It Too? Direct Realism and Objective Identity in Descartes." *Topoi*; special issue on "Direct Realism – Historical and Systematic Perspectives".

Soames, Scott (1987). "Direct Reference, Propositional Attitudes, and Semantic Content." *Philosophical Topics* 15(1), 47–87.

Stalnaker, Robert (2009). "What Is De Re Belief?" In Joseph Almog and Paolo Leonardi (eds.), *The Philosophy of David Kaplan*. Oxford & New York: Oxford University Press, 233–45.

Tachau, Katherine (1988). *Vision and Certitude in the Age of Ockham: Optics, Epistemology and the Foundation of Semantics 1250-1345*. *Vision and Certitude in the Age of Ockham*. Leiden: Brill.

Tweedale, Martin (2007). "Representation in Scholastic Epistemology." In Henrik Lagerlund (ed.), *Representation and Objects of Thought in Medieval Philosophy*. Aldershot and Burlington: Ashgate. 63–79.

Wettstein, Howard (1986). "Has Semantics Rested on a Mistake?" *Journal of Philosophy* 83, 185–209.

Wettstein, Howard (1990). "Frege-Russell Semantics?" *Dialectica* 44(1), 113–35.

Wiggins, David. 2001). *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.

Wilson, Margaret. 1978). *Descartes*. London: Routledge and Kegan Paul.

Wittgenstein, Ludwig (2009). *Philosophical Investigations*. (Edited by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte.) Revised 4t. Oxford: Blackwell.

Yolton, John (1984). *Perceptual Acquaintance from Descartes to Reid*. Oxford: Blackwell.

Carnapian Frameworks Revisited*

MATTI EKLUND

1. Introduction

In his (2021), Gabriel Broughton criticizes my work on Carnap on ontology and puts forward his own interpretation of what Carnap's external/internal distinction amounts to. I will here first argue that Broughton's main claims about my work are based on a misinterpretation, and then turn to some issues of broader interest. I will argue that Broughton's own, potentially interesting interpretation of Carnap's external/internal distinction does not work. And in light of the remarks on Broughton's discussion I will present a sharpened version of what I have earlier said about this distinction.

2. On Carnap's metaontology

Let me first, as background, go through what I have argued in earlier work regarding Carnap on ontology, and specifically the distinction between internal and external questions (Eklund 2009, 2013, 2016[1]). The focus has been slightly different in my different articles, but a common theme has concerned what Carnap's distinction between external and internal questions amounts to. My main points have, in brief, been the following.

Somehow or other, the distinction between internal and external questions is a distinction between questions internal and external to *frameworks*. So a basic question concerns what a framework is. Some Carnap commentators have taken

* Many thanks to Nils Franzén and to an anonymous referee (for a journal which in the end unfortunately decided against publishing this paper).

[1] Compare also my (2011) and (forthcoming), which are parallel but focus on Carnap's views on logic.

frameworks to be *languages*, or, better, *language-fragments*; others (or the same commentators in different contexts) have taken frameworks to be something more controversial, something which involves an interesting sort of *relativity* of the truth of a claim to a framework. On the former, *language pluralist* interpretation, an internal question becomes, in some way, simply a question internal to a language and an external question becomes, in some way, a question external to a language. The existence of frameworks becomes uncontroversial, but one may wonder how this is philosophically significant, and one can wonder what a question external to a language even might be. On the latter, *relativist* interpretation, Carnap is something more like a, well, relativist, holding that the truth of *claims* is relative to frameworks (where this is different from the trivial relativity of truth of *sentences* to languages). The claim that there are numbers may be true, and analytically so, relative to some framework, but have a different status relative to some other framework.

One contribution I sought to make is simply that of highlighting this choice point, and the fact that different interpreters of Carnap have made different choices at this point. But I also argued that the language pluralist interpretation is more plausible. Let me just quote my (2013) summary of the reasons for holding this:

> Carnap calls the frameworks *linguistic frameworks* (when the article was first published he used the label *frameworks of entities*, but this was changed for when the article was reprinted in *Meaning and Necessity*). Where Carnap introduces the notion of a framework, he says, "If someone wishes to speak in his language about a new kind of entities, he has to introduce a system of new ways of speaking, subject to new rules; we shall call this procedure the construction of a linguistic *framework* for the new entities in question". In other words: to speak in one's language about some entities one needs to introduce the expressions by means of which to do so. Later, Carnap uses "thing language" to denote a framework. This is clear evidence that for Carnap, frameworks are fragments of languages. Moreover, if Carnap were a relativist, one would expect him to emphasize that truths of the relevant kind are true *only relative to some framework*, but such formulations are entirely absent from his article; generally,

> Carnap treats frameworks as something straightforward. (Eklund 2013, 233–34)

I also noted that there are some reasons for caution about both interpretations mentioned, understood as general interpretations of what is going on in Carnap's discussion of external and internal questions. Critics of Carnap, such as Quine, have taken the external/internal distinction to be bound up with the analytic/synthetic distinction, and Carnap seemed to *agree* (Eklund 2013, 236). But, as I discussed, on neither of the two interpretations offered is there a tight connection between the distinctions.

In my articles, I have pushed back on interpretations of Carnap which impute to Carnap relativism or some other kind of reliance on an interesting form of relativity. Start with relativism. In my (2009) and (2013), I brought up André Gallois (1998), David Chalmers (2009), Ted Sider (2001) and Scott Soames (2009) as (sometimes) speaking of Carnap as a relativist. Chalmers (2009) speaks of Carnap as holding that "there are many different ontological frameworks, holding that different sorts of entities exist"—different entities exist according to different ontological frameworks (Chalmers 2009, 78). That is, for obvious reasons, very different from saying merely that in different frameworks, different sentences of the form "there are ____s" come out true. Turning to other forms of relativity, in my (2011) I criticized Warren Goldfarb and Thomas Ricketts' talk of what "can be made sense of only within a linguistic framework" (Goldfarb and Ricketts 1992, 69). If "linguistic framework" just means *language* then this talk just amounts to talk of what can only be made sense of within a language. This sounds rather more trivial than one may suspect Goldfarb and Ricketts intend. Don't we need a language when making sense of anything at all? Further, Goldfarb and Ricketts take Carnap to "reject language-transcendence." They take this to be a central feature of Carnap's philosophy. Again, I noted that there is a somewhat trivializing understanding of what they say: to reject language-transcendence could just be to emphasize that we must always employ some language or other.

3. Broughton on Eklund on Carnapian frameworks

Enter Gabriel Broughton's recent (2021). Broughton's article is a largely framed as a critical discussion of what I have had to say about Carnap. Broughton declares his intentions early on, saying "In this paper, I show that Eklund misreads Carnap, and I argue that this misreading obscures fundamental features of Carnap's philosophy" (Broughton 2021, 4098).

As I will get to later, there are some genuine, and potentially interesting, differences between Broughton's own preferred interpretation of Carnap and the interpretation that I have defended. But most of Broughton's discussion of my work is devoted to attacking a position that simply is not mine, and much of my discussion here will be aimed at getting those matters out of the way.

Describing my view, Broughton says, "A Carnapian framework, [Eklund] says, is just a natural language" (Broughton 2021, 4098). He thus ascribes to me the following thesis:

> *The natural language thesis*. A Carnapian framework is a natural language.

It follows from the natural language thesis that constructed formal languages are not frameworks. Opposing this thesis, Broughton says, inter alia,

> First, if we assume that frameworks are natural languages, then it is hard to make sense of a whole bunch of things that Carnap says in 'Empiricism, Semantics, and Ontology' (henceforth: ESO), including his ubiquitous references to *constructing* and *establishing* frameworks, his insistence that a crucial step in the formation of a framework is the introduction of certain *variables*, and his explicit focus on the *specialized* languages of science and semantics. (Broughton 2021, 4098)

and:

> … since Carnap says in ESO that a variety of sentences are analytic in this or that framework, Eklund's interpretation of frameworks as natural languages conflicts with Carnap's repeated insistence, both before and after ESO, that no natural language sentence is properly called analytic. (Broughton 2021, 4099)

He concludes, "[t]ogether, these considerations show that frameworks are not natural languages."[2] (In what follows, I will, like Broughton, refer to Carnap (1950) as ESO.)

Broughton is attacking a straw man. I do not subscribe to the natural language thesis. A first and main point to make is that the notion of a *natural language* plays no role what so ever in any of my main claims, summarized above. I speak generally about *languages*, and there is no obvious reason at all, given the nature of my claims, why the languages specifically would have to be natural languages. In my texts on Carnap, the important contrast is between the language pluralist interpretation according to which frameworks are languages, and a different, relativist interpretation according to which they are something which is bound up with relative truth in some interesting sense. Nowhere in my works do I say that frameworks are *natural* languages. Nor do I ever argue for such a claim. Any thesis of mine to the effect that frameworks are natural languages would be unstated, unargued, and unnecessary. These points alone should, to put it mildly, make one a bit skeptical of the view that I hold or have held such a thesis. Moreover, given Carnap's seemingly evident interest in and fondness for constructed languages, and the evident focus on such languages in ESO, any thesis to the effect that frameworks are natural languages would be surprising, not to say bizarre. And it would be all the more bizarre to maintain such a thesis without bothering to either state it or defend it.[3]

---

[2] I have here elided some considerations Broughton brings up in the longer passage I am quoting from, having to do with so-called pragmatic-external questions. I will introduce these considerations only later.

[3] I might add that in a blog post from January 28, 2021, André Carus (2021), one of the two authors of the (2020) *Stanford Encyclopedia* entry "Rudolf Carnap," brings up Broughton's article and describes Broughton as someone who "has now decided that enough is enough, and sprayed some serious ant killer on irruption of philosophical insect life." Generally, Carus says that "during the past few years a lot of rubbish has been circulating about Carnapian frameworks." (As examples of "rubbish," Carus mentions not only my work on Carnap but also that of David Chalmers (2009).) Turning to specifics, what Carus mentions Broughton as having shown, as against me, is that frameworks are not natural languages. Again, the thesis under attack simply is not mine.

Consider also Broughton's own summary of my main claims:

> An internal question—a question posed *within* a framework—is thus a question posed in a language. An external question, understood as a question about a matter of fact, would be a question posed in no language at all. No wonder Carnap found such questions unintelligible. On the other hand, Eklund suggests, the practical question of which language to speak seems perfectly intelligible. Again, just as Carnap suggested.
>
> If this reading is correct, then it refutes Quine's claim that the internal/external (I/E) distinction is bound up with the analytic/synthetic (A/S) distinction. In fact, on this reading, the I/E distinction does not seem to be bound up with much of anything that one might find problematic. The notion of a framework looks downright trivial. (Broughton 2021, 4098)

I basically find this a fair summary of my view, even if I will get to some complications below. But note that on Broughton's own summary of my view, any insistence on the frameworks being *natural* languages would be completely otiose. What reasonable work could "natural" even do, when inserted before the particular occurrences of "language"?

There is even an internal tension between different theses Broughton appears to ascribe to me. In the passage just quoted, he ascribes to me the view that internal questions are questions posed in languages, and external questions would *hence* be questions posed in no language at all. The "hence" is unstated but I take it to be conveyed by Broughton's "would." But if we take frameworks to be natural languages and only natural languages, the reasoning would seem to amount to: "An internal question is a question posed in a natural language; an external question would be a question posed in no language at all." There would be an obvious retort: couldn't an external question be asked in a non-natural, constructed language?[4]

[4] In the main text, I am concerned to show how Broughton misreads me. One question that arises is what explains Broughton's misreading. One possibility is that Broughton (to my mind somewhat reasonably) thinks it is so obvious that frameworks are languages of some kind that it cannot possibly be that obvious point I am making—and so he reinterprets me as

4. Broughton's reasoning

The natural language thesis would be quite startling given common knowledge of Carnap. Moreover, I neither state it nor argue for it, and it is unnecessary for my purposes. But of course, none of these points *conclusively* shows that I have not relied any such thesis in my work on Carnap. I could have surreptitiously relied on such a thesis. So let us take a look at the reasons Broughton adduces, and otherwise might have, for ascribing the thesis to me.

First, Broughton fastens on the fact that I use natural language examples when discussing Carnapian theses. In my discussions, I do keep using natural language examples when discussing frameworks and one may take this to be a reason for ascribing to me the natural language thesis. But the mere fact that I use natural language examples should not be accorded much weight: as Broughton himself notes, Carnap does too.[5] More importantly, already if it does not matter what kind of language is used, one can stick to natural language examples, which have the advantage of being familiar. Moreover, and more specifically, consider the following alternatives to the natural language thesis given which it is perfectly natural and reasonable to use natural language examples:

> *The permissive thesis*. Both natural and other languages are frameworks in Carnap's sense.
>
> *The indifference thesis*. Carnap's aims when drawing the external/internal distinction are such that it doesn't matter exactly which sorts of languages are at issue.

---

meaning something more specific, natural languages, where he speaks of "languages." But as described in the last section, there are various interpretations of Carnap which take him to invoke something relativism-like.

[5] In section 5 of his article, Broughton argues that Carnap's own use of such examples is compatible with rejection of the natural language thesis. I agree, but would disagree with the further claim that this is in tension with my interpretation—for the reason I do not ascribe the natural language thesis to Carnap.

Both these theses are compatible with Carnap's having independently held views, e.g., about the messiness of natural language, which led him to focus on constructed languages.

There are two slightly different versions of the indifference thesis. One (immodest) version claims that Carnap's overall outlook was such that he was indifferent to the question of what sorts of languages are at issue. Another (modest) version claims merely that for a general understanding of the internal/external distinction and its use in metaontology it does not matter whether natural or constructed languages are at issue. The modest version is compatible with the claim that Carnap for independent reasons, perhaps a desire to exclude natural languages due to their messiness, would only have counted constructed languages among frameworks.

Given either of these theses, the use of natural language examples is natural and justified. Given the permissive thesis, natural languages are some of the frameworks there are. Given the indifference thesis, it is a matter of indifference, as far as the external/internal distinction and its uses are concerned, whether natural languages are among the frameworks. Again, it makes sense to use natural language examples, for they do not do any harm and they do not require as much set-up.

Neither the permissive thesis nor the indifference thesis involves the bizarre claim that constructed languages would fail to count as frameworks. And return now to some central points Broughton brings up against me. In a passage already quoted, Broughton emphasizes Carnap's "ubiquitous references to *constructing* and *establishing* frameworks, his insistence that a crucial step in the formation of a framework is the introduction of certain variables, and his explicit focus on the *specialized* languages of science and semantics" (Broughton 2021, 4098), and the claim (which I will return to later) that Carnap held that no natural language sentence is analytic. These points are perfectly compatible with both the permissive thesis and the indifference thesis. All that they show is that constructed languages can be counted among the frameworks.

Either of the weaker theses would justify my use of natural language examples. But it is not even clear that the weaker theses are needed for what Broughton himself summarizes as

my main claims. Again, all I need is that frameworks are *languages*.[6] No further details about the status of these languages as natural or constructed are relevant given my aims.

In addition to focusing on my use of examples from natural language, Broughton adduces the following piece of evidence. It has to do with my talk of what language we "actually employ" and "actually use" in my (2009). I think those formulations of mine were somewhat unhelpful. But they do not indicate what Broughton seems to think they indicate. Here is the relevant bit from my paper, quoted by Broughton:

> If "framework" means language-fragment, the internal questions are those that concern what comes out true in the language we actually employ; pragmatic-external [questions] concern which language it is useful to employ; and factual-external questions are neither and thus by Carnap's lights make no sense. Here is an analogy. One can imagine three different debates, two of which are in order and one confused, that all can be brought under the heading "Is the tomato a fruit or a vegetable?" (1) Most straightforwardly, we can conceive of a debate over whether the [sentence] "the tomato is a fruit" is true as turning on what actually comes out true in our common language, English. When you and I discuss the matter, then you win if you say "the tomato is a fruit" and this sentence actually

[6] Here is a further reason why it is odd to ascribe the natural language thesis to me. In (2013), discussing Scott Soames, I quote Soames saying "[Carnap's] key thesis is that ontological questions are intelligible only within a scientific framework for describing the world. Such a framework is a formalized (or formalizable) language, with semantic rules interpreting its expressions, and assigning truth conditions to its sentences" (Soames 2009, 428, quoted in Eklund 2013, 235). Soames is here explicit that he holds that for Carnap a framework is a formalized or formalizable language. But when discussing this, I only discuss the fact that for Soames, a framework is a language (and notes that this seems incompatible with other things Soames goes on to say). If I subscribed to the natural language thesis, or even generally found it important that natural languages must be counted among frameworks, one would expect me to somehow mark disagreement here. The alternative would be to take me just to simply fail to notice the disagreement with Soames over this point. Thanks here to the anonymous referee I mentioned in the general acknowledgments.

> is what comes out true in our language. Taken thus, it is an internal question. (2) Somewhat less straightforwardly, perhaps, we can imagine a debate where the disputants are less concerned with what comes out true in English as actually spoken, but are concerned with whether it would be more pragmatically useful to speak a version of English just like English except for the possible difference that "the tomato is a fruit" comes out true there. Taken thus, the debate is over a pragmatic-external question. (3) Most obscurely, we can imagine two disputants who announce that they are not concerned with what comes out true in English—perhaps both agree that "the tomato is a fruit" is best English—and who further announce that they are not concerned with a pragmatic question of how we should speak. They announce that what they are concerned with is whether, in some language-independent sense, the tomato really is a fruit. If it is hard to wrap one's mind around what this would amount to, that is because these disputants would be seriously confused.[7]

Commenting on this, Broughton says:

> The first point that I want to make is just the one that I flagged above, viz. that Eklund takes Carnapian frameworks to be *natural* languages. He arguably suggests as much when he says that internal questions concern what comes out *true in the language we actually employ*, since we actually employ natural languages. But his commitment to this reading comes out even more clearly in the course of his discussion of the debates over "The tomato is a fruit" and "There are numbers." In the tomato example, Eklund tells us that the internal question concerns whether "The tomato is a fruit" comes out true *in English.* Meanwhile, the pragmatic-external question concerns whether it would be useful to speak an *English-like* language in which "The tomato is a fruit" comes out true. And similarly in the numbers dispute. I conclude that, in general, Eklund takes Carnapian frameworks to be natural languages or slight variations thereof. (Broughton 2021, 4103–04)

The fact that the example is from natural language is a feature that is irrelevant for the argument. To show this, let me

[7] Eklund 2009, 133. Quoted in Broughton 2021, 4103.

switch the example to one involving some formalism. Consider a sentence of the form "~(p & ~p)," of some given constructed language, and consider different sorts of disputes between a classical logician and a dialetheist concerning this sentence. First, there is a possible object-level dispute. One assertively utters this sentence; the other utters its negation and adduces evidence for it, and the dispute concerns whether that sentence, in the language they both employ, is true. If the language to which the sentence belongs is a constructed framework with explicitly laid down rules, that dispute can be easily settled. Second, while *using* that same sentence they can in fact be engaged in a dispute over whether, for pragmatic purposes, it would be best to use a (formal) language where this sentence comes out true. This would be an instance of metalinguistic negotiation, in Plunkett and Sundell's (2013) terms: non-metalinguistic sentences are used to issue conflicting metalinguistic recommendations. Third, the disputants are agreed both on what truth-value the sentence has in their common language (or their respective languages if they use different ones) and on pragmatic matters, but still have an attempted dispute over whether "it *really* is the case that ~(p & ~p)."

This is exactly the tripartite distinction I draw in the passage quoted. The distinction is obviously as applicable in the case of constructed languages as in the case of natural languages. Again to stress, one can certainly use a natural language example without thereby committing to the natural language thesis.

What then about the use of the "actually"? The use of the "actually" is there in order to distinguish one kind of use of a sentence from other kinds of uses that can be made of it. In the relevant use what matters is what comes out true in the language the disputants employ; and it is natural to use "actually" to emphasize the point.[8] In the revised formulation of

[8] In a footnote Broughton mentions the possibility of this alternative reading of the use of "actually" (Broughton 2021, 4103, fn. 6), which makes it odd to stress the use of "actually" to support his reading of me. (In the passage at issue, I speak of our "actual language" in the singular. It would be in line with Broughton's reading of me to say that on Carnap's view, internal questions can only be raised in one language: the one language

my point, the object-level dispute turns on what constructed language the logicians in fact—or "actually"—use.

5. The weaker theses

Broughton ascribes the natural language thesis to me. The natural language thesis is obviously false. Moreover, as I have shown, it is not reasonable to ascribe it to me. One may think that none of this need matter much in the grander scheme of things, if Broughton also showed that the weaker theses discussed in the last section are false. But first, as already stressed, I do not even need the weaker theses. Second, Broughton shows no such thing. Arguing against the natural language thesis, Broughton makes points such as the following:

> … ESO is filled with creation talk. We hear about *constructing* frameworks and *establishing* them. We hear about *introducing* expressions and *laying down* rules. None of this would be at home in a discussion of the properties of a natural language. What's more, Carnap says that a crucial step in the construction of a framework is the introduction of certain *variables*. Yet everyday English makes no use of variables. Carnap also makes frequent reference in ESO to *specialized* languages, specifically languages associated with the sciences and philosophical semantics.
>
> Carnap seems to be concerned less with ordinary English than with, as he puts it, the language (or, perhaps, the mere *calculus*) of mathematics, the language of physics, and so on. (Broughton 2021, 4105)

These are relevant points against the natural language thesis. But the fact that Carnap is so preoccupied with constructed languages can show nothing more than that constructed languages of a certain type are among what Carnap calls frameworks, and that Carnap finds these constructed languages to be of special interest. None of this speaks against either of the weaker theses.

---

we currently use. That would an interesting, but decidedly odd, interpretation of Carnap…)

As for it being an, as Carnap puts it, "essential" step in the construction of a framework to introduce variables, a main point to stress is that it is one thing to say, as Carnap does, that this step is crucial in the construction of a given framework and another to say that this step is crucial in the construction of any framework.[9]

Attention to the context where Carnap says this also shows that what is going on is that Carnap thinks, following Quine, that it is existential quantification in a formal language that carries ontological commitment. Recall here the modest version of what I called the indifference thesis. This view on existence talk and ontological commitment may provide a reason to focus exclusively on formal, constructed languages in discussions of ontology, but it is a view on existence talk that is separable from any appeal to an internal/external distinction.

Later in his discussion, Broughton appeals to the supposed fact that Carnap denied that sentences of natural languages are properly called analytic and notes that in ESO, Carnap "is perfectly happy to apply the term ["analytic"] to sentences formulated in a framework" (Broughton 2021, 4108). I am not as sure as Broughton seems to be that Carnap's considered view was that natural language sentences are never analytic. But however that may be, Broughton's argument here again at most shows that some sentences of some frameworks are not natural language sentences. This again is compatible with either of the weaker theses.[10]

---

[9] One may in principle question whether it is so obvious that English does not use variables. But let this pass.

[10] The remarks in the main text suffice as a response to what Broughton says about analyticity, but there is more to add. Broughton does adduce seemingly compelling evidence for the claim that for Carnap no natural language sentences are analytic. He quotes Carnap saying:

> the analytic-synthetic distinction can be drawn always and only with respect to *a language system*, i.e., a language organized according to explicitly formulated rules, not with respect to a historically given natural language (Carnap 1990, 432, quoted in Broughton 2021, 4108; Broughton's emphasis added).

As noted early on, I have stressed in earlier work that given my interpretation of Carnap, the internal/external distinction is not bound up with the analytic/synthetic distinction. Given that it at least seems that Carnap agrees with Quine that the two distinctions are closely connected, this is a potential problem for me. Broughton takes it to be a point in favor of his view that he respects Carnap's view on the connection, saying "While it's always possible that Carnap somehow misunderstood his own views, surely, all else being equal, we should prefer an interpretation that avoids this result" (Broughton 2021, 4119). The idea is that given his proposal there is the following connection: if frameworks are formal languages and formal languages are characterized in part by semantic rules, then formal languages will contain analytic sentences, corresponding somehow to these semantic rules. But I am doubtful regarding the truth of this conditional claim.

One may think no further evidence is needed. This is as explicit as it gets. But other things that Carnap says blur the picture. The very same paper that Broughton quotes from begins as follows:

> It must be emphasized that the concept of analyticity has an exact definition only in the case of a language system, namely a system of semantical rules, not in the case of an ordinary language, because in the latter the words have no clearly defined meaning. (Carnap 1990, 427)

This is different. Here Carnap is not saying that the analytic-synthetic distinction *cannot be drawn* with respect to ordinary language, but only that analyticity does not have an "exact definition" with respect to ordinary language. I think the evidence regarding Carnap and the analyticity of ordinary language sentences is equivocal. Moreover, the whole of Carnap (1955) is an apparently constructive attempt to make sense of synonymy—and hence, by Carnap's lights, analyticity—in natural languages.

What is more, some things Broughton himself says are in tension with holding that for Carnap no natural language sentences are analytic. In connection with defending (I/E), Broughton, as I will get to, allows that some natural language sentences can be straightforwardly translated into what by Broughton's lights are framework sentences. But then these natural language sentences can be said to be governed by semantic rules corresponding to framework sentences, and generally have semantic features corresponding to the framework sentences, including analyticity.

Broughton says that if, like me, one denies that the two distinctions are bound up with each other, then one holds that Carnap *misunderstood* his own views, given that Carnap held that the distinctions are related. But there are other possibilities. For example, one possibility is that Carnap simply held further views given which the views are bound up with each other. And in fact, what Carnap says is:

> Quine does not acknowledge the distinction which I emphasize above, because according to his general conception there are no sharp boundary lines between logical and factual truth, between questions of meaning and questions of fact, between the acceptance of a language structure and the acceptance of an assertion formulated in the language. (Carnap 1950, 215, fn. 5)

Here Carnap appears to equate acceptance of the analytic/synthetic distinction (that there is a "sharp boundary" between "logical" and "factual" truth), with accepting that there is a distinction between "acceptance of a language structure and the acceptance of an assertion formulated in the language." But it seems that one can agree with Quine regarding the analytic/synthetic distinction even while holding that it is one thing to decide to speak a language and another to accept given assertions formulated in that language. To put things more plainly: Quine took his rejection of the analytic/synthetic distinction to allow him to play fast and loose with the distinction between languages and theories, and Carnap seemed to agree, but there is no reason to go along with this.

6. Broughton's positive proposal

Let me now turn to Broughton's own positive proposal regarding what Carnap's internal/external distinction amounts to. I will both discuss the proposal in its own right, and how the positive proposal relates to my understanding of Carnap and Broughton's criticisms of me. The proposal is this:

> (I/E) An internal question is a question that can be straightforwardly translated as the question whether φ, where φ is a sentence of some framework S, and φ is understood to have the meaning assigned to it by the semantical rules of S. An external

> question is a question that is not an internal question. (Broughton 2021, 4118)

A framework is, in turn, a formal language "endowed with a syntax, a semantics, and a confirmation theory" (Broughton 2021, 4099). Broughton further thinks that for Carnap many (questions corresponding to) sentences of natural language—all sentences such that it is too unclear what they mean—fail the test for being internal in this sense, and so fall on the side of external questions. As formulated, Broughton's proposal of course straightforwardly entails that frameworks are not natural languages.

I have expressed concerns about how Broughton discusses my work on Carnap. But even if Broughton's criticisms of me are misguided, it could be that his own positive proposal is a better proposal than what I have presented.

Before assessing Broughton's proposal, let me stress that Broughton's positive view actually is congenial to much of what I want to say. On Broughton's view as on mine, Carnap has no truck with relativism, and the talk of frameworks itself is relatively straightforward and uncontroversial. Moreover, note that an internal question can for Broughton be one formulated in natural language. It is fully consistent with Broughton's proposal to use natural language examples of internal questions. All that is needed is that it be possible to translate the natural language sentences into sentences of a suitable formal language. (Although Broughton adds, reasonably, that Carnap thought that due to the messiness of natural languages such translation will seldom or never be determinately correct (Broughton 2021, 4117).[11]) Moreover, there is nothing in Broughton's proposal that is in tension with the alternatives to the natural language thesis that I discussed earlier.

It is independently plausible that for Carnap, translatability into a framework sentence is a necessary condition for (cognitive) meaningfulness. But then the translatability condition in Broughton's (I/E) just amounts to a meaningfulness condition.

---

[11] Broughton does think that for Carnap what users of natural language mean in the sense of speaker meaning may be more determinate.

That said, I am not persuaded by Broughton's proposal. My concerns are straightforward. For Broughton, any question that is deficient in meaning in such a way that it cannot be translated into a sentence of a framework—i.e., for Broughton, of a suitable constructed language—is an external question. But Carnap is clear that he has in mind something much more specific by "external" than Broughton allows: he has in mind a certain kind of philosophical question. In his Intellectual Autobiography (1963)—which Broughton himself centrally appeals to—he says:

> In accord with my old principle of tolerance, I proposed to admit any forms of expression as soon as sufficient logical rules for their use are given. If a philosopher asks a question like "are there natural numbers?", he means it as a question so-to-speak outside the given language, raised for the purpose of examining the admissibility of such a language. Therefore I called philosophical questions of existence of this kind external questions. (Carnap 1963, 66)

Remarks like this leave no doubt that Carnap meant something rather specific by "external" in such a way that not every question that fails to be internal in Broughton's sense is external. Earlier, in ESO, Carnap says, "From the internal questions we must clearly distinguish external questions, i.e., philosophical questions concerning the existence or reality of the total system of the new entities" (Carnap 1950, 214)—and on Broughton's interpretation, the "i.e." should have been an "e.g." Carnap, I might add, throughout only uses philosophical questions about the existence or reality of some new entities as examples of external questions. By itself that may be meagre evidence against Broughton's proposal: Carnap could be using these specific examples just because ontology happens to be the topic at hand. But together with Carnap's explicit statements about what he takes external questions to be, these facts about what examples Carnap uses provide further evidence against Broughton's interpretation. Carnap's external questions all have a certain distinctive philosophical flavor; the class of vague or unclear questions posed in natural language is certainly bigger than that.

A central feature of Broughton's own proposal is that it treats the internal/external distinction as exhaustive. All

questions are either internal or external: any question that does not meet the conditions for being internal counts as external. The labels "internal" and "external" do of course suggest that the distinction is exhaustive: a question is either *inside* or *outside*, whatever exactly this means. But I see no reason to think that the distinction in fact must be exhaustive or that Carnap's discussion indicates that it has to be. And if the class of external questions is narrow in the way I have argued, it would be odd to take the internal/external distinction to be exhaustive. For what it is worth, Carnap's examples of internal questions indicate that they too always in some way concern existence. I do not see that anything I have said commits me to a particular stance on the issue. More importantly, I do not see that I need to take a stand on this. The important point for me is that what an internal question is internal to, and what an external question is external to, is a language, and it is of less importance whether all questions internal to languages count as "internal" and all questions external to languages count as "external."

Even if Carnap's distinction is not intended as exhaustive, a modified version of his proposal still could work. Broughton might say:

> (I/E*) An internal question is a question that can be straightforwardly translated as the question whether φ, where φ is a sentence of some framework S, and φ is understood to have the meaning assigned to it by the semantical rules of S. An external question is a *certain type of* purported non-internal question concerned with the existence of the entities postulated by the framework.

As I will get to in the next section, one may want to add a restriction regarding what counts as an internal question parallel to that added regarding external questions. I will not get into further discussion of (I/E*). For reasons noted above in connection with (I/E), it would not be problematic for me to accept that thesis. And I do not see that anything in my general outlook on Carnap commits me to thinking that the internal/external distinction is exhaustive, so I have no problem with the modification involved in (I/E*). Of course, the "certain type" is vague and anyone defending (I/E*) may wish to say more about that clause.

7. "Internal" and "external" revisited

I have defended my interpretation of Carnap against what I take to be Broughton's chief objections, which involves gratuitously imputing to me the natural language thesis, and I have criticized (I/E), Broughton's alternative interpretation of Carnap's distinction between internal and external questions. But let me end on a more constructive note, and by making a concession to Broughton. In addition to other points he seeks to make, Broughton criticizes my take on "external" versus "internal" for being unduly simple. For example, if an internal question is simply one internal to a language and an external question is one where one tries to stand outside language and so asks no question at all, where do the pragmatic-external questions fit in?[12]) More specifically, can't a supposed pragmatic-external question be raised perfectly well in a suitable language and would it not then be internal? At least in natural language one can certainly ask things like: ought we to use this language or that?

I believe Broughton is entirely right to raise questions regarding this aspect of my discussion. Before returning to what I have earlier said, let me first focus on how best to describe the distinction between internal, pragmatic-external and factual-external questions within the overall picture that I present.

Let me first note that perhaps one ought not to expect very much precision. Carnap's labels "internal" and "external" may be evocative and useful—indeed, the popularity of appeal to the distinction may have to do with how evocative the labels are—but Carnap did not offer necessary and sufficient conditions for falling in either category. This omission may be perfectly justifiable: a distinction can be useful despite failing to be completely clear and sharp. Moreover, the specific labels "pragmatic-external" and "factual-external" are from me. While the distinction is there in Carnap, it is less emphasized and Carnap does not even try to label the distinction. There is then some reason to suspect that problems may arise when

[12] Broughton 2021, p. 4098–99. This point about pragmatic-external questions is the part from Broughton's summary of his criticisms of me that I elided earlier.

one tries to be more careful about what that distinction amounts to.

However, that said, there actually are some helpful things to say.

I have already criticized the assumption that the internal-external distinction is exhaustive. Given that the distinction is not exhaustive, the question "where do the pragmatic questions fit in" does not have the same bite. They could form a separate category. But there still remains the question: why are these pragmatic questions not a subspecies of internal questions?

One way to respond to this question is to say that not all questions in some sense internal to the kinds of languages at issue (whether these are natural languages, constructed languages, or both) are internal in Carnap's sense. Just as all external questions in Carnap's sense are intended as having to do with existence, all internal questions have to do with existence. Pragmatic-external questions are not internal because they are not themselves existence questions.

Getting more specific, I find the following passage in ESO very helpful:

> On the other hand, the external questions of the reality of physical space and physical time are pseudo-questions. A question like: "Are there (really) space-time points?" is ambiguous. It may be meant as an internal question; then the affirmative answer is, of course, analytic and trivial. Or it may be meant in the external sense: "Shall we introduce such and such forms into our language?"; in this case it is not a theoretical but a practical question, a matter of decision rather than assertion, and hence the proposed formulation would be misleading. Or finally, it may be meant in the following sense: "Are our experiences such that the use of the linguistic forms in question will be expedient and fruitful?" This is a theoretical question of a factual, empirical nature. But it concerns a matter of degree; therefore a formulation in the form "real or not?" would be inadequate. (Carnap 1950, 213)

What Carnap speaks of as "ambiguity" seems to be the fact that a given form of words may be used to convey different things. The form of words "are there space-time points?" can, first, simply be used to ask whether there are space-time

points, as in general "are there *F*s?" can be used to ask whether there are *F*s. But the form of words may also be used in different ways. It can be used to raise a practical question—the pragmatic-external question. The form of words is then used to convey something other than convey what the sentence semantically expresses. The label "external" is rather apt because the questioner seeks to view the language from the outside, even if, of course, the sentence "should we speak a language in which 'there are space-time points' comes out true?" would express the same thing, and count as internal. The form of words can also be used to ask the "factual, empirical" question Carnap mentions at the end of this passage.

Finally, although he does not say so in the passage just quoted, I take Carnap to hold that there are philosophers who would be apt to use the same form of words to try to ask a different question, one that is not the internal question, not the practical one, and not the factual, empirical one about efficiency, but is a philosophical question about the reality of the entities in question. It is this question that Carnap takes to be a chimera.

As should be clear, Carnap actually distinguishes between *four* different kinds of questions. There is the internal question, the practical (pragmatic-external) question, the "theoretical question of a factual, empirical nature"—and then the kind of (confused) external question that purports to be a genuine theoretical question. I wonder if there is not yet another problem for Broughton here. I do not see why the "theoretical question of a factual, empirical nature" could not be an internal question in the sense of Broughton's Carnap. But Carnap evidently does not class such a question as internal.

I think that my reasoning in the "the tomato is a fruit" case very well captures the sort of issue that Carnap's internal/external distinction concerns.[13] There too we have the one and the same form of words that may be used to raise different issues. There is the straightforward issue of whether "the tomato is a fruit" is true in the language used by the speaker (this is what I spoke of as the language actually employed). There is the practical—pragmatic-external—question

[13] Leaving aside the fact that Carnap focused on existence questions, of course.

of which language to speak (and the corresponding question about efficiency). And there is the confused, supposedly deep philosophical question.

While both the practical and the confused question may be called "external," they are "external" in quite different ways. The confused question is external in that it aims to be a question raised in no language at all and in that sense external to all language. A pragmatic-external question is external not in that sense but in the sense that it serves to ask questions *about* languages, *assessing* them. Both kinds of questions can be called external, but they are external in different ways.

8. Concluding remarks

Broughton ascribes to me the view that for Carnap, frameworks are exclusively natural languages. This is a misunderstanding on Broughton's part. Broughton's discussion of his own positive thesis regarding Carnap's external/internal distinction is better, and his positive ideas hold more promise, but I have explained why this positive thesis should be rejected. Finally, I turned to the constructive task of, within my general picture, accounting for Carnap's distinction between internal questions, pragmatic-external, and (supposed) factual-external questions. Along the way, I have discussed whether the internal/external distinction is exhaustive, and I have noted that in Carnap there is a distinction between four kinds of questions.

Uppsala University

References

Broughton, Gabriel (2021). "Carnapian Frameworks." *Synthese* 199, 4097–4126.

Carnap, Rudolf (1950). "Empiricism, Semantics and Ontology." *Revue Internationale de Philosophie* 4, 20–40. Reprinted with minor changes in Carnap (1956), 205–221.

Carnap, Rudolf (1955). "Meaning and Synonymy in Natural Languages." *Philosophical Studies* 6, 33–47. Reprinted in Carnap (1956), 233–247.

Carnap, Rudolf (1956). *Meaning and Necessity: A Study in Semantics and Modal Logic*, enlarged edition. Chicago: University of Chicago Press.

Carnap, Rudolf (1963). "Intellectual Autobiography." In P.A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*. La Salle, Ill.: Open Court, 1–83.

Carnap, Rudolf (1990). "Quine on Analyticity." In Richard Creath (ed.), *Dear Carnap, Dear Van: The Quine-Carnap Correspondence and Related Work*. Berkeley: University of California Press, 427–32.

Carus, André (2021). "Frameworks Vindicated", blog post, https://awcarus.com/2021/01/frameworks-vindicated/. Retrieved July 31, 2021.

Chalmers, David (2009). "Ontological Anti-Realism." In Chalmers, Manley, and Wasserman (2009), 77–129.

Chalmers, David, David Manley, and Ryan Wasserman (eds.) (2009). *Metametaphysics*. Oxford: Oxford University Press.

Eklund, Matti (2009). "Carnap and Ontological Pluralism." In Chalmers, Manley, and Wasserman (2009), 130–156.

Eklund, Matti (2011). "Multitude, Tolerance and Language-Transcendence." *Synthese* 187, 833–47.

Eklund, Matti (2013). "Carnap's Metaontology." *Noûs* 47, 229–49.

Eklund, Matti (2016). "Carnap's Legacy for the Contemporary Metaontological Debate." In Stephan Blatti and Sandra Lapointe (eds.), *Ontology After Carnap*. Oxford: Oxford University Press, 165–189.

Eklund, Matti (forthcoming). "Carnap, Language Pluralism, and Rationality." In Darren Bradley (ed.), *Philosophical Methodology After Carnap*.

Gallois, André (1998). "Does Ontology Rest on a Mistake?" *Proceedings of the Aristotelian Society, Suppl. Vol. 72,* 263–83.

Goldfarb, Warren, and Thomas Ricketts (1992). "Carnap and the Philosophy of Mathematics." In David Bell and Wilhelm Vossenkuhl (eds.), *Wissenschaft und Subjektivität. Science and Subjectivity: The Vienna Circle and Twentieth-Century Philosophy*. Berlin: Akademie-Verlag, 61–78.

Leitgeb, Hannes, and André Carus (2021). "Rudolf Carnap." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), https://plato.stanford.edu/archives/sum2021/entries/carnap/.

Plunkett, David, and Tim Sundell (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13, 1–37.

Quine, Willard Van Orman (1948). "On What There Is." *Review of Metaphysics* 2, 21–38.

Sider, Theodore (2001). *Four-Dimensionalism*. Oxford: Oxford University Press.

Soames, Scott (2009). "Ontology, Analyticity and Meaning: The Quine-Carnap Dispute." In Chalmers, Manley, and Wasserman (2009), 424–443.

The Semantics of Common Nouns and the Nature of Semantics

JOSEPH ALMOG & ANDREA BIANCHI

This paper should be read as a sequel, more than fifty years on, to Putnam's breakthrough "Is semantics possible?" (1970). This early piece is less celebrated than Putnam's later ones and those of Donnellan, Kripke, and Kaplan, turnabout papers that mark *the referential turn* against the Frege-Carnap classical model of, (1), the semantics of (proper and common) *nouns* in particular and, (2), the form of *semantic theory* in general. We believe that "Is semantics possible?" hides its light under a bushel; it is deeply illuminating both on the specific topic of noun-reference and on the more general question of what kind of science semantics is. We would like to revisit both issues half a century on.

In the space of a few pages, Putnam manages to touch what seems to him two related topics, that is, (1), the semantics of common nouns in natural languages, and, (2), the question that gives the paper its title, "Is semantics possible?". How are (1) and (2) connected for Putnam? For the sake of argument, Putnam accepts Quine's then most influential general skepticism about semantics as an empirical scientific theory, say on the model of chemistry or biology. The paradigm developments in formal semantics (ubiquitous in that inventive decade, the Sixties, just before Putnam wrote) followed the structure of abstract *model theories* of formal languages. In a word (playing on a formulation made famous

later by Partee) semantics as practiced appeared more like a branch of mathematics (viz., model theory, algebra) than of the natural sciences.[1] In the model theories of formal languages, the lexical (atomic) base is treated *schematically* and the focus is rather on the (sentential) connectives. This gives us *recursive* semantics, where a semantic rule is associated with each syntactic rule that generates new forms out of ingredient inputs. But if, says Putnam following Quine, all that we can do is to assimilate natural languages to formal ones and offer schematic model theories for them, the prospects of semantics as a natural science are dim. However, continues Putnam, not all hope is lost. We *can* and thus *should* investigate the lexical base of natural languages, their nouns, adjectives, verbs, adverbs, etc. As we do so, an *empirical* theory of genuinely natural-historical phenomena, *natural languages*, emerges that even Quine might see as scientific. The naturalization of semantics starts by de-schematizing it and attending to its lexical base.

Putnam goes on to do just this. He investigates the prospects for a theory of common nouns in natural languages.[2] Through this investigation, he comes to see various things. To begin with, the classical reductive and reference-free Frege-Carnap *predicative* semantics of common nouns is in error: common nouns are *referential*. Furthermore, now at the higher level of engaging with Quine's challenge, his discoveries in the test case of common nouns impart a host of morals about how to make a natural science out of semantics.

We acknowledge Putnam's methodology and in particular the symbiotic connection between the specific topic of what the semantics of common nouns is and the general one of what a semantics of natural languages should be. In this paper, we see ourselves as amplifying his points regarding (1) and (2). As for (1), there emerges a uniform referential seman-

---

[1] See Montague 1970a, 1970b, and 1973, and Partee 1979. Montague's famous title "English as a formal language" conveys the gist of the method if the phrase "as a formal language" is read in a strong way, as indeed Montague intended: the English fragment is *reduced* to a formal (higher-order) language, with its own logical syntax cum model theory.

[2] Later, he comments on verbs like "grow" and adjectives like "red" (1975, 244). Here we focus on common nouns, although we believe that our remarks can be extended to all categorematic words.

tics for (common) nouns. A host of familiar problems, in our view due to intrusions from metaphysics, are now dissolved. This leads towards the end of the paper to launching reflections related to (2), the proper treatment of natural language semantics in general.

1. Some guidelines for a semantics of common nouns

The following are our fundamental guidelines for a semantics of common nouns:

> A. *The uniformity of nouns I*: All nouns, proper ("Aristotle") and common ("water," "tiger"), are to be treated uniformly. From a semantic point of view, all nouns function in the same way.
>
> B. *The uniformity of nouns II*: Any of the aforementioned nouns, e.g., "tiger," has the same semantic function wherever it occurs in a sentence.[3]
>
> C. *The uniformity of nouns III*: The sole semantic function of nouns is to refer.
>
> D. *No predicative reduction*: No noun is to be reduced to a predicate (open sentence).
>
> E. *No extensional reduction*: The semantic value of a noun is not an extension (in a model, world).[4]
>
> F. *No intensional-modal reduction*: The semantic value of a noun is not a modal intension.

(A)–(C) insist that nouns are to be treated uniformly. By means of (A), we exclude the unprincipled reductions practiced by the classical revisionist logical-form tradition. For example, Russell allowed some proper nouns, but not others, to be reduced to predicates (descriptions); in sophisticated later variations, *all* proper nouns were admitted as non-predicative, though common nouns were still reduced, as a matter of standard formal symbolization, to predicates. Ac-

---

[3] Davidson called such a feature *semantic innocence* (1968–9, 108). See also Barwise and Perry 1981.

[4] By this, of course, we do not mean that there is no set collecting things that a common noun is *true of*. But one should not take this set to be *semantically* related to the noun.

cording to (A), we cannot treat "Aristotle" referentially while we treat "tiger" predicatively: either all are reduced to predicates (as indeed suggested by the strict classical model) or else all refer. We view Putnam as accepting thesis (A) all the way down.

We should like it noted that thesis (A) does not yet settle whether nouns refer: we may well let them all be reduced to mechanisms of predication as indeed urged by the logical tradition when driven by generality and elegance (as in the work of Quine on the elimination of all "singular terms"). In a similar vein, (B), according to which a given noun, proper or common, functions semantically in an invariant way, without shifts created by this or that embedding context, is a formal uniformity thesis (for a given noun in all its occurrences), but not yet a thesis telling us what the function of the noun is (in all these occurrences).

Our third thesis, (C), is that common nouns, like proper nouns, refer to worldly entities: just as "Aristotle" refers to the man Aristotle, "tiger" refers to the animal kind tigers.

Thesis (C), according to which the semantic function of nouns as such is to refer, was not developed in full by the aforementioned quartet of pioneers of the referential turn. They certainly made it clear that proper nouns refer and are not predicative but have left it open whether common nouns do so.[5]

To understand this thesis requires a two-step move. The first is to separate it from a host of non-semantic, frankly *metaphysical*, theses that have blurred our understanding of the semantics proper. Then, once the metaphysical intruders are out of the way, we need to focus on the primal semantic relation, *reference*.

2. Semantics vs. the intrusion of metaphysical doctrines

We shall point to three major intrusions of metaphysical doctrines that have clouded the possibility of referential semantics for all common nouns (there may well be others). The

[5] Donnellan (1983), for example, was somewhat skeptical about the referentiality of common nouns.

first involves the injection of defining predicates ("characteristic marks") in the actual world while stating what a common noun such as "tiger" stands for. This leads to subordination of the noun's semantics to the metaphysics of what is referred to by it. It thus occurs that true predications about the kind of animals precede and determine the reference of the noun. The second is an amplification of the first, this time with involvement of *modal* predications, alleged necessary truths projected across possible worlds about the kind, prematurely infesting the semantics. The third concerns the idea that the existence and identity conditions of some kinds but not others (e.g., the *artifactual* kind pencils but not the *natural* kind tigers) depend on our linguistic activities and this difference must be reflected in the very semantics of the common nouns we use to refer to them.

2.1 The intrusion of actual true predications

Metaphysical questions about the existence and identity of the entity referred to need to be separated from a discussion of the *semantic* relation (*reference*) between the noun and the entity referred to. We should investigate noun-semantics without speculating about the referred entity's metaphysics and investigate the referred entity's metaphysics without speculating about noun-semantics; in a nutshell, substantial metaphysics without noun-semantics and noun-semantics without substantial metaphysics.

Observe the independence in the seemingly simpler case of proper nouns and the individuals they refer to. We may ask the metaphysical question (whether about the *ontology* and nature of reality or in terms of *modal* issues of trans-world identity) whether Aristotle had (of necessity, of his essence or his nature) to be generated by a particular sperm and egg and, thus, at a particular time in history. To do so requires no specific doctrine about how the name "Aristotle," the demonstrative "he" (uttered pointing to Aristotle) or a description, definite or indefinite, "the (an) author of the *Nicomachean Ethics*" relates semantically to the philosopher. Get the man proper, by whatever means, and you can ask a question about him and his existence and identity conditions, and if you so will, project it even modally, to how he *must* have originated.

In like manner, we may ask the metaphysical question about the species (kind) of tigers, whether it had to originate by a reproductive mechanism in a certain ur-group with a given DNA at a certain period in history, e.g., only so many million years ago, and on planet Earth. We may ponder all this whether we use the single word "tiger," the Latin (now scientific) expression "Felis Tigris," the description "my favorite feline species," or the complex demonstrative "that kind of animal" (uttered pointing to Shere Khan).

Similar observations on independence from semantic doctrines about nouns apply to the metaphysics of trans-world relations dissected by Kripke in the case of individuals, e.g., that Nixon of this world and Nixon of that (any) other world must share the *same* O-relation ("same origin"), and in the case of kinds by Putnam, according to whom some ingredients in the real world and some ingredients in another world are of the same kind (if and) only if they bear the theoretical *same* L-relation (same chemical structure, same DNA, etc.). These are all claims of metaphysics, concerned with what makes an entity (individual or kind) the one it is (across worlds). They are not questions about the semantics of nouns.[6]

So much for the independence of metaphysical questions. In the reverse direction, semantic questions about *how* nouns refer to entities should not be mixed with questions about the properties of the entities proper. This, again, is quite clear in the case of proper nouns. The noun "Aristotle" refers to Aristotle and this is no observation, in metaphysics, about the *truth* of any *predicate* applying to that man. The latter type of question concerns the *satisfaction* relation, obtaining between Aristotle and a compound predicate, e.g., "is identical to Aristotle," "is the man originating in gametes X," or "is the author of the *Nicomachean Ethics*," all true of Aristotle, though the last only contingently, the middle one of necessity but not on grounds of logic alone, and only the first necessarily and on *logical* grounds.

---

[6] These questions could be raised in a formal language (e.g., in the quantified modal language of Kripke 1963) even if all singular terms and specifically all individual constants were eliminated. Variables have values but do not refer.

Quite apart from any metaphysical doctrine, the predicates and the noun relate in different ways to their semantic values. Predicates are said to have an extension or denotation or designation (the last is Carnap's 1947 term), which is the *set* of items satisfying them.[7] If the extension is down to a singleton set, it is still a *set* that is the extension and the same is true even if it is this fixed singleton set that serves as the sole ("rigid") extension of the predicate across all possible worlds. We can say then that the extension is *modally* rigid but in spite of the spellbinding effect this phrase has had in philosophy this still just means: a certain set has been coming up consistently as the extension of the predicate across worlds. In this respect whether the set is a singleton, a doubleton ("is a square root of four") or an infinite set ("is a prime") does not alter the fact that the *set,* not an individual or a kind, has served as the *extension* (not as the *referent*) for the predicate in all worlds. On the other hand, the proper nouns "Aristotle" and "Omega" refer to particular individuals, an ancient Greek and the first infinite ordinal, regardless of the satisfaction of any predicate by the man or the number. The question of what the noun refers to (Aristotle, Omega) is prior to any predication of that man or that number, let alone modalized (necessary, essentialist) predications.

The pattern we have just observed with proper nouns recurs with common nouns. If we consider the trio of kind-describing predicates "is Obama's favorite kind of animal," "is the kind of animal with DNA D" and "is the same kind as Shere Khan's actual infima species," we encounter predicates whose extension is, respectively, contingently correlated to the referent of "tiger" (the kind tigers), necessarily so related, and logically necessarily so related.[8] The extensions of the predicates, rigid or not, are not (are never!) the referent of the word "tiger," the kind tigers. If we now approach the kind, as reductive metaphysics has urged, by means of trans-world *extensions* of the kind (from which we *construct* the kind), the

---

[7] We ignore here predicative locutions such as "is an ordinal (a set)," which may have, not in a model but in the absolute universe V, a correlate too large to be comprehended as a set.

[8] We assume it a logical validity of the pertinent modal logic "If actually P, then necessarily actually P."

difference between predicate-designation and noun-reference recurs. If we consider predicates of individual animals not of the kind proper, e.g., "is an animal that is a member of Obama's favorite kind", "is an animal with DNA D" and "is an animal of the same kind as Shere Khan's actual kind," we get as extensions three sets. None of these sets of animals is the kind (which is never a set). In a nutshell, proper and common nouns that *refer* do not have (rigid) extensions and predicates that have extensions (rigidly or not) do not refer.

2.2 Modalizing extensions

The foregoing discussion should simply dissolve a problem deemed grave in the transition from proper to common nouns in modalized semantics. It is often said that the key fact concerning the semantics of a proper noun such as "Aristotle" is that it *rigidly designates* Aristotle. When we want to extend this allegedly key notion from proper to common nouns, a crisis strikes: the common noun "tiger" seems to designate different sets (of tigers) in different possible worlds. Thus "tiger" would be a non-rigid designator.

The problem is bogus and could have been seen to be such by either considering a case such as "prime" where the alleged extension (designation) would be rigid or assuming for the sake of the argument a metaphysics, like Spinoza's and other modal determinists', in which there are no counterfactual worlds, the way the world *is* is the only way it *might* have been. In such a set up only one set, the actual set of tigers, would be designated by "tiger." But in both cases, be it that of "prime" or "tiger," this would still, rigid extension and all, get things wrong because these sets are not what "prime" and "tiger" refer to. The sets are still *assembled* only by way of *satisfaction* by each of their members of a certain key predicate: they depend on truths such as *two is a prime*, *three is a prime*, *five is a prime* and *Shere Khan is a tiger*, *Tony is a tiger*, *Tigger is a tiger*, etc. This is the way in which we may assemble the rigid extension (in our modal deterministic set up) of the predicate "is an animal with stripes etc." or "is a number divisible only by itself and one". These two sets are the (rigid) *extensions* (*designations*) of the two predicates but they are not the *referents* of the two nouns. We simply evalu-

ate the predicate world by world and in each get an extension, a certain set of individuals. It may then turn out that one and the same set is obtained throughout the worlds. But whether it is the same set in all worlds or not, the referent of "tiger" and "prime" is another thing.

This is exactly as it is with proper nouns and their alleged rigid "designations." The word "Aristotle" *refers* to Aristotle; it has no rigid extension (designation) because it has no extension (designation) to begin with. What we want to say rather is that the individual the noun "Aristotle" refers to, Aristotle himself, is the entity that is relevant to evaluations of modal predications, be it in a primitive modal language ("might not have been a philosopher," "is necessarily human") or in the possible world alternative vocabulary ("is a philosopher (human) in *w*").

The notion of designation, which applies to predicates, is indeed *world-relative*. A special case of it is *rigid* designation, wherein the same designation keeps coming up throughout the spectrum of worlds. In contrast, the notion of reference is not world-relative at all: the neologism "refers in *w*" has been an error from the outset confusing *model theory* (which does define extension (designation) at a model (world)) and *semantics* and the *mundane* relation of referring.

"'Aristotle' refers to Aristotle" is absolutely either true or false, period. In this case, it is true and the referent, Aristotle himself, is the only thing that matters for modal predication. Should the claim be false, as in "'Aristotle' refers to Plato," it is false once and for all. It is for this simple reason that Kripke's (1972, 24, 156–8) insight both about the empty proper noun "Vulcan" and the empty common noun "unicorn" is so important: if the noun is actually empty, if it *fails* to refer, that is it; there is no redeeming of the failure in other worlds. On the other hand, a predicate such as "is an animal with one horn looking like a horse," whose extension is empty in the real world, could of course have a non-empty extension in alternative worlds. In a similar vein, notice that a predicate such as "is an even prime that is not two" has an empty extension in all worlds because as we keep evaluating, no satisfier ever comes up. This leaves the compound predicative expression meaningful. It is a case very different from that of

the noun "unicorn," which fails to refer to anything whatsoever. It is truly empty of any semantic value.

To sum up: nouns do not designate (have extensions), they refer (or fail to refer). This much is prior to any truth of a predication about the referent. It is predicates that designate, rigidly or not, depending now on the satisfaction of the predicate by candidate individuals/kinds across worlds. The intuition that in saying "Trump (tigers) might have lost the battle" we assess "might have lost the battle" of the *actual* referent, the individual Trump and the kind tigers, is correct: of that referent we consider a modal predication or a predication holding of it in an alternative world *w*. Nowhere is there any question of reassessing the *reference* of "Trump" ("tiger") in another world.

2.3 Different kind of kind, different semantics?

Let us come now to the third metaphysical intrusion into the semantics of common nouns. It has often been suggested that *artifactual* kinds such as that of pencils metaphysically differ from *natural* kinds such as that of tigers.[9] E.g., at the level of *individual essentialism* it has been claimed that whereas an individual tiger is of necessity a tiger, a pencil might not be of necessity a pencil. More critical yet, at the level of *kind essentialism* it has been pointed out that to be a member of the kind pencils something needs to have a certain function (and a certain appearance) perhaps due to stipulations (intentions) of the *designer* of the artifact. In contrast, to be a member of the kind tigers something must have a certain DNA and descend from tigers and this is beyond the control of any designer. And now, in a final step of *semantic reflection*, this purported metaphysical difference between the kinds is projected in the semantics of the corresponding nouns: "tiger" would be governed by the deep structure kind-essentialist condition but "pencil" would have a classical descriptive meaning given by a functional or appearance level description.[10]

---

[9] See, e.g., Schwartz 1978 and 1980.

[10] Another, non-equivalent but to many related, way of making the point is that to be a member of the kind tigers one must bear the same X-relation to some actual paradigm tigers but no such theoretical relation is

Two claims are made here, one belonging to metaphysics about the kinds proper, the other about the reputedly reflective semantics of the nouns. Our purpose in this semantic paper is not to discuss the metaphysics of kinds (this independence is indeed one of our points), but we note in passing that the metaphysical claim about the kinds is anything but obvious.

First, the distinction between artifactual and natural kinds often seems to be hastily overdrawn. The fact that a certain kind depends for its existence on the actions of thinking (human) beings is not sufficient, because such beings produce distinct kinds of products, e.g., distinct types of shadows or sweat or noises or liquids that are unique to those kinds (and to individuals of the kind: a human being's shadow of the human walking could not have existed if humans were not walking and our walking-shadow can only exist if we produce it).

If Aristotle is the inspiration in separating natural and artifactual products, one may point to a key distinction in terms of *intentions* and goals and why not *final causes* governing the artifacts but not the natural products. This is a common philosophical distinction, e.g., between two isomorphic rock-made objects, a rock naturally shaped by an erupting volcano and a rock shaped by a sculptor who carved an ashtray out of it. But the distinction may presuppose a dubious metaphysics of humanly uncaused original acts, of some freely chosen actions outside the frame of natural laws and totally segregated inside the heads of intenders outside space and time causation. The common philosophical presumption of a sort of *actus originarius* outside the causal framework whereby the designer is creating a new kind out of nothing by means of an inner template seems to be an abstraction from the process of handling concrete materials (e.g., the way in which ashtrays are fashioned from hardened lava materials).

---

at work in the case of membership in the kind pencils. And now, having made the point about the kinds proper, it is urged that the noun "tiger" expresses as its meaning the deep structural relation to a paradigm. For some criticisms of the *semantic reflection* step, see Bianchi 2022.

Secondly, it seems that a leap from epistemic considerations to metaphysical assertions is at work. What has been called the *reference fixing* or *identifying* description by means of which the kind is introduced to an immaculate audience, thus giving the audience epistemic access to *which* kind is in question, takes the metaphysical role of *defining* the kind tout court. This conversion seems incorrect, for often artifactual kinds turn out not to abide by the original designer's identification; the kind has a life of its own and it mutates so as to be made now, e.g., of new alloys not previously available or of synthetic rubber or genetically engineered materials. Likewise what was intended by the designer for purposes of religious worship may find a use/function in saving the lives of the tribe's babies. No original stipulation can control forever what happens to the kind. Just like a natural living kind, it evolves and mutates, exactly as natural-historical individuals do. The original designer is not the metaphysical controller.

Finally, as pointed out by Putnam and especially by Burge in a series of landmark papers in the late Seventies, even if it were true that metaphysically some constitutive condition governs any possible pencil or sofa, this is far from having this kind of condition available in the head of a common user of the word "pencil" or "sofa." The competent user may be just as much in the dark about iPhones and gaskets as she is about elms and beeches.

The foregoing is meant to note *en passant* that any idea that the artifactual kinds proper are somewhat controlled by the recipes we have in our head is dubious. But now, to return to the main, semantic, point of this paper, let us just assume that the artifactual kinds are indeed metaphysically different from the natural kinds. We may even assume that to be a pencil is essentially to look like normal pencils and be used like them, whereas to be a tiger it is neither necessary nor sufficient to look like normal tigers and act like them. So, let us, for the purpose of the discussion, admit two categories of kinds, those governed by deep structure and natural-historical conditions and those controlled by designers' definitions. To add to the menu, we may further consider *mathematical* kinds which, at least in some views in the philosophy of mathematics, are given by *a priori* definitions and could not turn out any different from how they have been defined. How does

this affect our account of the semantics of the nouns we use to refer to those kinds?

The simple answer we give is: it does not affect it at all. Again, the case of individual essences and proper noun reference to the individuals having these essences should offer the clear simple model. Let us suppose that the noun "Nixon" refers to the human being Nixon who, as part of his true metaphysics, had to originate in gametes X. At the same time, the noun "Shmixon" refers to a certain *person* related to the human being. But, as Locke observed long ago, *person* is a "forensic." Many who would let the person Shmixon be individuated by his memories or other psychological profiles would surely skip over the sperm and egg origin (just as those who are focused on the human being Nixon originating in that zygote do not make the memories criterial). And of course, we can introduce a succession of such nouns for forensically defined items, all the way to a pure Cartesian ego ("Dixon"), who may not need a body at all to exist. So there; the entities Nixon, Shmixon, Dixon etc. surely differ in their existence and identity conditions. Nonetheless, the nouns refer to these three (and other such) in the same way, directly and not by means of the satisfaction of any condition. It is true that each of the three referents satisfies a different structural condition but it is not true that what makes the entities the *referents* of the three nouns is satisfaction of such conditions. Semantic reflection is false: the difference in the metaphysical profiles of the entities is not reflected in a difference in the type of semantics for the three nouns. The nouns refer to the three entities, each noun having as its sole semantic function to refer. The entities referred to are of course substantially different metaphysically. In like manner, it may well be that what constitutes a biological vs. forensic (or artifactual) vs. mathematical kind involves different types of conditions. This purported difference in metaphysical profiles of the kinds is not reflected in a difference in the type of semantics for the related common nouns: each of them has as its sole semantic function to refer to a kind.

A potent example by which we may encapsulate this separation between semantics and metaphysics is Putnam's own famous case of the word "jade" (1975, 241). According to Putnam, as a matter of actual historical fact the word refers to

two different substances, jadeite and nephrite. Many options have been tried against this example. One idea is that the word "jade" is after all synonymous with a description detailing the surface features shared by (pieces of) jadeite and nephrite. Another option, at the other end, is that there are two words "jade" (as there might be two words "bank") each referring directly to a separate substance. A third option, not to be confused with the first, is that the word "jade" refers to one kind only, the kind (pieces of) jade, membership in which requires being a piece of either jadeite or nephrite (wherein satisfaction of the surface description is neither sufficient nor necessary).

We need not immediately make the choice of the correct resolution but we do need to exclude some incorrect options. The word "jade" is not synonymous with a surface description. Indeed, the user of the word may use it without having even that surface description "in the head," simply by receiving it from fellow users (who may or may not have a ready description to provide). In receiving the word, the new user goes on to *refer* to whatever her predecessors did without any guarantee that she will be as informed about the appearance of jade.

Should we say that there are two words "jade," each referring to its own chemical kind or should we say we have a single word that refers to the kind jade, membership in which involves being a piece of either jadeite or nephrite? We note that such questions recur both with proper and common nouns, wherein a single surface appearance can be received by the user carrying more than one meaning (referent). Thus the word "Aristotle" names various Greek men (and we may well suppose some of them look alike). In Putnam's case of the Twin Earth use of "water," which may well arise in two different ecologies on Earth, we again have one word, or two homonymous words *loaded* with two different substances, made to have a similar qualitative appearance.

To develop a stance on such cases we need to take our last step and understand the semantic relation of *reference*.

3. The fulcrum of semantics: user's (back-)reference

At the beginning of this paper, we read Putnam as reorienting semantics to focus on lexical items, in particular common nouns. Once we refocused on semantic investigations in this way, we observed, also following Putnam's inspiration, that such words as simple (proper and common) nouns *refer* rather than *denote* (*designate*). This is in sharp contrast with the long *reductionist* tradition(s) emanating from Frege, Russell, Quine, Carnap and Montague of those who held that they are in fact ("disguised") compound terms. This semantic reductionism sought out a dual semantic theory with a separation of meaning and something else, a semantic X-factor, so that, (1), meaning *determines* X (in a world, in a context). Furthermore, (2), this X-factor, the denotation or extension, is a *subsidiary* semantic value of the expression. Finally, coming to cognition, (3), what the user grasps or has in mind in using the expression is the primary semantic value, the meaning.

We saw that Putnam's reorientation of semantics towards lexical items came hand in hand with a reorientation concerning the fundamental semantic relations. Indeed, the reorientation revealed that there is a *unitary* such semantic relation, not a duality of meaning and denotation. This unitary relation we called *reference*. In contrast to denotation, reference does not run from an immaculate word to the object (kind) but rather in the opposite direction, from the object (kind) to the user; the referent is loaded into the word the user receives. Furthermore and related to this, semantics essentially involves the receiver mentioned, the user of words. Upon reception of a given word, the user *acts* with it to refer back to whatever object the word was already loaded with. Our semantics is one of (*back-*)*referring users* and *their uses*.

This much unites most modern referential theories. But differences emerge when one tries to reflect on what this unitary semantic relation of (back-)reference is. Often the differences simmer over test cases in which we witness a split over *which* objectual candidates the user is (back-)referring to. This is not a dispute about whether it is *reference* (e.g., as opposed to *denotation*) that is taking place, it is a dispute among referentialists over precisely *what* relation semantic reference is.

In speaking of the (user's) *semantic referent* of a noun we mean just that: the referent as semantically relevant (indeed, as the noun's sole semantic value). We must take care here not to read into our terminology the popular *theoretical* distinction introduced by Kripke (1977) between *semantic* reference and *speaker's* reference. The distinction has become standard nomenclature, as if theoretically innocuous. We shall not criticize it here but nor shall we use it.[11] Rather, we introduce our own terminology, *user's semantic* (*back*-) *reference*.

Our terminology is meant to record two key facts that could be missed by someone attending to Kripke's way of cutting the pie. We think it is essential to *natural* language that it is *users* (*agents*) who refer, using words as instruments. For us at the heart of semantics is the question of what the *user* refers to, by using a given word on an occasion of use. So the allusion to the user's actions is key.

4. The nature of semantics

Let us take stock and ponder what we have learned by attending to Putnam's game-changing paper.

We see a double-edged message. The first is *intra*-semantic**:** the discovery concerns internally the semantics of common nouns. The issue here is what kind of semantic values should be assigned to them by the semantic theory. The second is *meta*-semantic, as it concerns the very character of semantics as a science: what is the domain of semantics and what other investigations (pre- and post-semantic) need be separated.

The two levels are related for Putnam. They were already related in the history of the subject, in the days of Frege, Wittgenstein's *Tractatus*, Church, Carnap, Katz, Montague, Lewis and a host of other modern semantic theorists. We might think of the framework as inspired by Frege and brought into a modern form by way of the intension/extension systematic account offered by Carnap in his aptly called *Meaning and Necessity* (1947).

Within this framework, the crux of semantics is the *meanings* assigned to words in stage 1 of the semantics. We may

[11] For two somewhat divergent critical discussions of Kripke's distinction, see Almog 2012 and Bianchi 2019.

call the meanings *senses* or *intensions* or *concepts* (in the common, non-Fregean, sense of the word). Given such a meaning assignment and given a factual parameter (a model, a possible world, the real world), a derivative value is determined, e.g., the planet Venus, the set of planets (in a world, model etc.). The meanings operate at two critical levels: (1) they determine the worldly-extensions, external values that go into a calculation of truth values (Is Venus in the set of planets?); (2) they are *internalized* by competent language users of the words (grasping the meanings is what it takes to *understand* sentences, prior to any extension/truth value determination), and what determines translations between languages. Thus, they are the fundamental materials of semantics. This intra-semantic claim has consequences at the meta-semantic level: semantics is a self-sustaining science with words and compounds *already* endowed with meanings in stage 1 *before* we move to the next, evaluational, stage 2 of seeking the *post*-semantic and fact-dependent extensions, be they objects, sets or truth values associated with the linguistic media. The world enters this picture at a later stage, not until post-semantic stage 2, when we need to compute post-semantic extensional information. By this time, the language proper is fully semantically functional, expressing meaningful sentences, allowing translations and ready for understanding by the competent speakers, from whom we demand grasp of the basic meanings and ability to compose them using their syntactic competence. What is important to notice is that here there is a trade-off between (i) the self-sufficient internality of the science of semantics and (ii) the demotion of the real world from being a key determiner (it merely plays a post-semantic role as an extension-provider).

Putnam taught us to reject the double-edged thesis, both inside semantics and at the meta-semantic level.

His famous Twin Earth thought-experiments urge upon us the *impossibility result* that the meaning (at least of words like "water") cannot be both (1) what determines the worldly extension and (2) what is known by the competent speaker. Putnam produces cases where what the speaker knows (has "in the head") is simply not sufficient to determine the worldly extension.

Read literally, Putnam confronts us with a choice: (A) keep the meanings accessible to the speaker's head and give up their role as determiners of the worldly extensions or (B) acknowledge that the determination of extension operates in a quite different way and thus give up on the idea that meanings are available transparently to competent users. However, we prefer to read Putnam as questioning *both* (A) and (B), and to see the reflections on common nouns he offers as a verifying *test case*. The postulation of *meanings* is the original sin. Meanings have gone by the board and are not missed; they do not determine extensions, but, just as much, they are not what is known by competent speakers.

Thus, Putnam rejects the generalization of meanings-semantics *first*, then evaluation at many indices. He points to the need to ask for a reconfigured *meaning-free* and thoroughly *referential* semantics. But how on Earth did the words get their reference?

Now the world comes in not as a post-semantic evaluation point (where we look at a whole spectrum of possible worlds only one of which is "real"). The world that comes into play for Putnam is only the real world and its web of connections and it comes *prior* to semantics: we explain the very *possibility* of semantics by looking at the *origin* of our uses. The world, by its actions of dubbing, word generation, word transfer and causation of speakers to use words, determines the semantics of our uses, all the way down as in natural science. In this way, we isolate preconditions for semantics. There is a prior stage in which real world materials have (i) to *exist* and (ii) to be appropriately *linked* into a world-wide-web of *connected structure*. Real world processes made it the case that "Nixon" and "tiger" have a semantics (they refer to Nixon and to the kind tigers, respectively). In contrast, "Vulcan" and "unicorn" do not have a semantics, because there is no individual and no kind that has been connected to the users of the two words.

It is at this level of background facts of existence and connections that epistemological puzzles get resolved. How can a true identity sentence be informative, i.e., what *internal meaning* will make it so? How can we determine from inside the head the difference between "Neptune," which does refer, and "Vulcan," which does not, what meaning would do this

job? How can we determine from inside the head, without a causal background of connections to the users, whether they speak of Smith or Jones when they utter "Smith is raking the leaves"?

Putnam directs us to semantically deflating answers. In traditional semantics, we are looking for the key under the lamppost of meanings, when we should look at the dark side of the street; no meanings can answer world-involving questions. The answers do not lie in the head, they lie in the real world pre-semantic processes that determine what words are, on an occasion of use, connected with.[12]

University of Turku
University of Parma

References

Almog, J. (2012). "Referential Uses and the Foundations of Direct Reference." In J. Almog and P. Leonardi (eds.), *Having in Mind: The Philosophy of Keith Donnellan*. New York: Oxford University Press, 176–184.

Barwise, J. and J. Perry (1981). "Semantic Innocence and Uncompromising Situations." *Midwest Studies in Philosophy* 6, 387–403.

Bianchi, A. (2019). "Speaker's Reference, Semantic Reference, and the Gricean Project. Some Notes from a Non-Believer." *Croatian Journal of Philosophy* 19, 423–448.

Bianchi, A. (2022). "Kind Terms and Semantic Uniformity." *Philosophia* 50, 7–17.

Carnap, R. (1947). *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: The University of Chicago Press.

Davidson, D. (1968–9). "On Saying That." *Synthese* 19, 130–146. Reprinted in D. Davidson, *Inquiries Into Truth and Interpretation*. Oxford: Clarendon Press 1984, 93–108. (Page numbers given relate to the latter volume.)

Donnellan, K. (1983). "Kripke and Putnam on Natural Kind Terms." In C. Ginet and S. Shoemaker (eds.), *Knowledge and Mind: Philosophical Es-*

---

[12] We would like to thank Panu Raatikainen for his invitation to contribute to this volume and Paolo Leonardi, Ernesto Napoli, and Gabriel Sandu for their comments to previous drafts. We dedicate this work to the memory of Hilary Putnam and Saul Kripke, who were generous and inspiring to both of us in different ways over many years.

says. New York: Oxford University Press, 84–104. Reprinted in K. Donnellan, *Essays on Reference, Language, and Mind* (edited by J. Almog and P. Leonardi). New York: Oxford University Press 2012, 179–203.

Kripke, S.A. (1963). "Semantical Considerations on Modal Logic." *Acta Philosophica Fennica* 16, 83–94.

Kripke, S.A. (1972). "Naming and Necessity." In D. Davidson and G. Harman (eds.), *Semantics of Natural Language*. Dordrecht: Reidel, 253–355, 763–769. Reprinted with a new introduction as *Naming and Necessity*. Oxford: Blackwell 1980. (Page numbers given relate to the latter volume.)

Kripke, S.A. (1977). "Speaker's Reference and Semantic Reference." *Midwest Studies in Philosophy* 2, 255–276.

Montague, R. (1970a). "English as a Formal Language." In B. Visentini (ed.), *Linguaggi nella società e nella tecnica*. Milano: Edizioni di Comunità, 189–223. Reprinted in R. Montague, *Formal Philosophy* (edited by R.H. Thomason). New Haven: Yale University Press 1974, 188–221.

Montague, R. (1970b). "Universal Grammar." *Theoria* 36, 373–398. Reprinted in R. Montague, *Formal Philosophy* (edited by R.H. Thomason). New Haven: Yale University Press 1974, 222–246.

Montague, R. (1973). "The Proper Treatment of Quantification in Ordinary English." In J. Hintikka, J. Moravcsik, and P. Suppes (eds.), *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*. Dordrecht: Reidel, 221–242. Reprinted in R. Montague, *Formal Philosophy* (edited by R.H. Thomason). New Haven: Yale University Press 1974, 247–270.

Partee, B.H. (1979). "Semantics – Mathematics or Psychology?" In R. Bäuerle, U. Egli, and A. von Stechow (eds.), *Semantics from Different Points of View*. Berlin: Springer Verlag, 1–14.

Putnam, H. (1970). "Is Semantics Possible?" *Metaphilosophy* 1, 187–201. Reprinted in H. Putnam, *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press 1975, 139–152.

Putnam, H. (1975). "The Meaning of 'Meaning'." In K. Gunderson (ed.), *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science 7*. Minneapolis: University of Minnesota Press, 131–193. Reprinted in H. Putnam, *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press 1975, 215–271. (Page numbers given relate to the latter volume.)

Schwartz, S.P. (1978). "Putnam on Artifacts." *Philosophical Review* 87, 566–574.

Schwartz, S.P. (1980). "Natural Kinds and Nominal Kinds." *Mind* 89, 182–195.

The Fallacies of the New Theory of Reference: Some Afterthoughts[1]

GABRIEL SANDU

1. Introduction

The *New Theory of Reference* is a view according to which there is a subclass of expressions in natural language which are genuine naming devices in that they function as rigid designators, that is, they designate the same object in all possible worlds in which that object exists. The view has been mainly developed by Saul Kripke in his *Naming and Necessity* (Kripke is explicit though about not being after a theory) but elements from it may be found already in the work of Ruth Barcan Marcus in her talk "Modalities in Intensional Languages," delivered in 1962 at a session of the *Boston Colloquium for the Philosophy of Science*. In this paper I will reassess some of the claims made in Hintikka and Sandu, "The Fallacies of the New Theory of Reference" (1995). In that paper we denied the need for a class of basic expressions which function as rigid designators and claimed that the rigidity of those expressions can be expressed by using quantifiers. In the present paper I will qualify some of these assertions.

2. Kripke: Naming and Necessity

Kripke's famous lectures on *Naming and Necessity* (NN) were given in Princeton 1970, then published *verbatim* in Davidson and Harman (1972), and finally published by Harvard University Press as a book in 1980. The latter contains an Introduction in which Kripke tells us that he reached the main ideas of the monograph around 1963–64 based on his earlier

[1] I am greatly indebted to Joseph Almog for suggestions to improve the paper.

work in the model theory of modal logic. The earlier work refers to Kripke (1963), an article in which he developed a model-theoretical semantics for a first-order modal predicate language (with no individual constants but only individual variables). At the beginning of that article, we are told that:

> The authors closest to the present theory appear to be Hintikka and Kanger. The present treatment of quantification, however, is unique as far as I know, although it derives some inspiration from acquaintance with the very different methods of Prior and Hintikka. (Kripke 1963, 83, fn. 1.)

What is unique about Kripke's treatment of quantification in Kripke (1963)? It is the quantificational structure it imposes on a set of possible worlds (and the corresponding accessibility relation). It is such that:

- The Tarski-type notion of satisfaction of a formula is now generalized to a possible world, an interpretation of the non-logical vocabulary and an assignment to its free variables.
- Every possible world is endowed with its own domain of individuals which is the range of the quantifiers occurring in the formula (obeying the constraint: if an object exists in a possible world w, and w' is a distinct world accessible from w, then that object exists also in w'.)
- The individual assigned to a free variable does not depend on a possible world but is picked up, once and for all, from the union of the domains of the possible worlds ("rigid" interpretation of free variables.)

This semantic interpretation renders valid Leibniz's law of identity $\forall x \forall y(x = y \rightarrow Nec\ x = y)$. The language, however, does not contain individual constants and thus the semantic interpretation does not tell us anything about them, even less so about the interpretation of names in natural language.

In the Introduction to *NN* (1980) Kripke tells us that little by little (1963–64) he came to be convinced that names in natural language also function as rigid designators and that the necessity of identities holds for them too. The point about rigidity that Kripke emphasizes is that "we have a direct intuition of the rigidity of names, exhibited in our understanding

of the truth conditions of particular sentences." (Kripke 1980, 14) There are two kinds of such sentences that Kripke considers.

One of them consists of *simple sentence* like:

(i) Aristotle was fond of dogs.

There is general agreement that (i) is true if and only if a certain philosopher we call "Aristotle" was fond of dogs. But for Kripke our understanding of (i) requires more: we have to be able to recognize "the conditions under which a *counterfactual course of history*, resembling the actual course in some respects but not in others, would be correctly (partially) described by (i)" (Kripke 1980, 6). And that happens if and only if the same aforementioned man would have been fond of dogs, had the situation obtained." (Ibid.)

The other kind of sentences Kripke considers are *counterfactual sentences*:

> In the monograph I argued that the truth conditions of 'It might have been the case that Aristotle was fond of dogs' conform to the rigidity theory: no proof that some other person other than Aristotle might have been both fond of dogs and the greatest philosopher of antiquity is relevant to the truth of the quoted statement. (Kripke 1980, 12–13)

The intuition behind the two kinds of sentences considered by Kripke is that once a proper name, say "Nixon," names a particular person, it would continue to do so in all counterfactual scenarios in which that person exists. That led Kripke to develop his doctrine of proper names as rigid designators. As he tells us in the Introduction:

> … I imagined a hypothetical formal language in which a rigid designator 'a' is introduced with the ceremony, 'Let 'a' (rigidly) denote the unique object that has property F, when talking about any situation, actual or counterfactual'. It seems clearly that if a speaker did introduce a designator into language that way, then in virtue of his very linguistic act, he would be in a position to say 'I know that Fa', but nevertheless 'Fa' would express a contingent truth (provided that F is not an essential property of the unique object that possesses it.) First, this showed that epistemic questions should be separated from

> questions of necessity and contingency, and that to fix a reference is not to give a synonym (NN, 14).

Kripke's connection between names and modal questions in natural language was an important insight and it is difficult not agree with him that one of the main reasons for Russell's proposal of a theory of names (the so-called *descriptive theory*) incompatible with our intuitions of rigidity was his failure to consider modal questions.

3. The fallacies of the new theory of reference

In Sandu and Hintikka (1995) we argued that, contrary to what the proponents of the New Theory of Reference, including Kripke, hold, there is neither class of expressions (singular terms) which function as rigid designators nor primitive semantic phenomena of rigidity in natural language. We made our point by using, not the modalities of necessity and possibility, as Marcus and Kripke did, but epistemic notions like knowledge and belief whose logic Hintikka had analyzed in his *Knowledge and Belief* (1962). Let us shortly recall the basic steps.

In a first step, we rehearsed the well-known distinction between *de dicto* vs *de re* knowledge which seems to require two uses of certain singular terms (definite descriptions). Here is one of the examples we used.

In the *de dicto* case, someone, say a, may know something, e.g., that b is S, abbreviated by $S(b)$, of whoever is or may be referred to by the singular term "b." For instance, Stefan may know something about Marie Antoinette's lover, whoever he might have been, for instance that he was not French. We represent such knowledge in the logical notation by "$K_a S(b)$," where "b" stands for the description "Marie Antoinette's lover." The model-theoretic import of the truth of "$K_a S(b)$" is that in all the scenarios compatible with what a knows, it is the case that $S(b)$. But given that Stefan does not know who the gentleman in question is, the term "b" ("Marie Antoinette's lover") will pick out different individuals in the different scenarios compatible with everything Stefan knows.

On the other hand, in the *de re* case, a may know something about the individual who in fact is b, without knowing that he is b. For instance, Stefan may know some fact or other

about Count von Fersen, who in fact was Marie Antoinette's lover, even if Stefan does not know this fact about him. The decisive step is to observe that the truth of the knowledge statement requires the phrase "Marie Antoinette's lover" to pick out the same gentleman (viz. Count von Fersen) in all the scenarios admitted by Stefan's knowledge. In general, knowledge "of the individual who in fact is b" cannot be expressed by a statement of the form "$K_a S(b)$" unless "b" picks out the same individual in all the scenarios compatible with what a knows. In such case the term "b" designates whatever it designates necessarily, and it might seem that, in order to express *de re* knowledge, we must have at our disposal "rigid designators" referring to whatever they refer to necessarily. Furthermore, this rigid reference cannot be mediated by any contingent definite description. For such a description can always in principle refer to distinct individuals in different possible scenarios. We took the protagonists of the New Theory of Reference to identify their rigidly referential singular terms with proper names. However, this is not the strategy we endorsed.

In a second step, we expressed the "rigidity" of a definite description as a particular kind of *de re* modal attitude. We then observed that the same technique can be applied to express the rigidity of proper names. Finally, we pointed out, rehearsing some of Hintikka's earlier arguments in Hintikka (1969), that for quantifiers to perform this job, they must be interpreted referentially (objectually), which in turn presupposes a mechanism of cross-identification of individuals as denizens of various possible situations (worlds). We exemplified all these claims using a toy logical language. We recall again the main stages.

In a first stage, we consider sentences of the form:

(1) $\exists x N S(x)$

(2) $\exists x K_a S(x)$

where "N" stands for the necessity operator and "K_a" for the epistemic operator "a knows that." On the referential interpretation of quantifiers, the truth of (2) in a possible world w requires that there be an individual in w which belongs to the extension of S in all epistemic a-alternatives to w. Similar

truth conditions can be formulated for (1). Thus in both of these sentences, one is saying that something is true of one and the same individual in a range of different possibilities. In (1), the relevant possibilia are all the states of affairs or courses of events that are being considered possible. In (2) they are all the possibilities left open by what a knows.

In a second stage we express the *de re* interpretation of our earlier examples, using variants of (1) and (2). We first consider the *de re* interpretation of "$K_aS(b)$" in which a knows something about the individual who in fact is b, without knowing that he is b. This is rendered in our logical notation by:

(3) $\exists x(x = b \wedge K_aS(x))$;

and the analogue *de re* interpretation of "$NS(b)$" is expressed by:

(4) $\exists x(x = b \wedge NS(x))$.

Neither (3) nor (4) requires that "b" ("Marie Antoinette's lover") pick up the same individual in all the relevant alternatives, but only that the individual which is the actual referent of "b" belong to the extension of "S" in all these alternatives. In other words, the truth of both (3) and (4) is consistent with "b" picking up different individuals in various possible scenarios.

In a final stage, we consider the particular *de re* interpretation of "$K_aS(b)$" according to which a also knows who b is:

(5) $\exists xK_a(x = b \wedge S(x))$.

The corresponding *de re* interpretation of "$NS(b)$" is similarly expressed by:

(6) $\exists xN(x = b \wedge S(x))$.

We can actually abstract from the claim "$S(b)$" and express the rigidity of "b" simply by:

(7) $\exists xK_a(b = x)$.

And likewise, we can express the rigidity of "b" in alethic contexts by:

(8) $\exists xN(x = b)$.

The truth of (5)–(8) forces "b" to refer to one and the same individual in all the relevant alternative worlds, including the actual one (we ignore here some problems concerning the non-existence of individuals). In other words, "b" acts as a "rigid designator," something that we also alternatively expressed as "a knows who b is."

Finally, we realized that the same "rigidifying" strategy works independently of whether "b" stands for names or definite descriptions. This led us to conclude that there is no need to assume any class of singular terms in natural language which act as "rigid designators." We expressed this in the paper in the following way:

> …as soon as we have quantifiers at our disposal, we do not need any other kind of direct representability. In sum, the right slogan of modal logicians should be: We do it with quantifiers. And this dispensability seems to invalidate all arguments for the need of rigid designators or anything remotely like them in natural or formal languages. (Hintikka and Sandu 1995, 252–253.)

Let me emphasize two points about our approach in the paper. One of them, which we often repeated, was that the strategy of imposing the rigidity of a singular term by an outside quantifier works because we interpreted quantifiers "referentially." That is, quantified formulae are interpreted with respect to a possible world and an assignment, the latter assigning individuals (from the joint domain of discourse) to the free variables which occur in the corresponding open subformulae. Thus, recalling our earlier example (we assume here that if an individual exists in a possible world, then it also exists in all its relevant alternatives):

> (9) $\exists x K_a(b = x)$ is true in a possible world w with respect to the assignment g if and only if there is an individual $\beta \in dom(w)$ such that $b = x$ is true in every a-alternative w' with respect to the assignment $g\left(\frac{x}{\beta}\right)$ if and only if there is an individual $\beta \in dom(w)$ such that the individual who is the semantic value of "b" in w' is β. (Ibid., 249)

We observe that the interpretation of the constant "b" inherits its "rigidity" (i.e., constancy of its semantic value in all the a-alternative possible worlds) from the "rigidity" of the varia-

ble "x" induced by the referential interpretation of the existential quantifier "$\exists x$" which binds it. There is nothing new here, as we also acknowledged in the paper: this line of reasoning has been countenanced much earlier by Kripke himself in Kripke (1963). In that paper he treats quantifiers in alethic contexts in a referential way; and in Kripke (1976), he makes the distinction between *de re* and *de dicto* interpretations of definite descriptions and observes that there is way of expressing *de re* belief by using quantifiers (ibid., 374):

(10) $\exists x(x = b^*$ and Jones believes that x is an airdale).

Here "b^*" is a definite description. (10) is essentially the same as our example (3) above. Kripke actually uses this formulation for a language which does not contain explicit scope indicators. One would think that such a language is a fragment of our natural language, and thereby does not contain quantifiers and variables. I will say something about this below.

The second point to be emphasized about our approach in the paper is that the strategy we followed to impose rigidity as a particular kind of *de re* epistemic attitude works, obviously, only for modal contexts. I will return to this issue below.

4. Criteria of cross-identification

Although the strategy we followed to impose rigidity in our paper relies on a referential treatment of quantifiers in modal contexts due to Kripke himself, it is not the strategy he finally endorses with respect to the rigidity of names. I will say something about this in the next section. For now, let me shortly comment on another major philosophical disagreement between his treatment and ours, technicalities aside. Its source lies in the requirement of an individual to be a denizen of several possible worlds. Or, we thought in the paper, echoing some of Hintikka's earlier work, such a requirement presupposes criteria of cross-identification:

> As a slogan, we may perhaps put it, quantifying in presupposes that criteria of cross-identification have been given. These criteria cannot themselves be expressed by quantifiers. For in order to do so, we must be able to compare the denizens of any two

> scenarios ("possible worlds") for identity. (Hintikka and Sandu 1995, 249)
>
> Bound variables do not, in any literal sense, refer to anything at all. The rigidity and directness they exhibit is not a matter of reference but of criteria of cross identity. (Ibid., 253)

The requirement poses no problem for Kripke for whom possible worlds "are little more than the miniworlds of school probabilities blown large" (Kripke 1980, 18). We recall in this context Kripke's well-known example with two dice being thrown. There are 36 possible outcomes, that is, 36 states of the dice that Kripke takes to be 36 possible worlds. One of them is the state (die A, 6; die B, 5); another one is (die A, 5; die B, 6), etc. These possible worlds are abstract, not complex physical entities, and there is no need for some further criteria to compare e.g., die A, 6 in the first world with dies A, 5 in the second world. All in all, for Kripke, philosophical questions like "Which die is that?" simply do not make sense, for, as he observes, the states of the dices are simply *given*. (Ibid., 17)

The requirement of criteria of cross-identification in my paper with Hintikka, on the other side, amplified some of Hintikka's ideas in the late sixties (which finally go back to Carnap's "individual concepts") and was motivated by the way we understand the truth-conditions of certain belief sentences. It is well known from the rich industry of epistemic puzzles that singular terms in such sentences do not seem to behave "rigidly" and this, in turn, seems to have something to do with the modes of identification of individuals. Whether the latter is somehow related to the question of the substitutivity of names in belief contexts, as Hintikka thought in his earlier work, a view we endorsed in the paper, is a difficult matter, one which I will not deal with here. My main concern is more modest, viz., to reassess the claim we made in the paper to the effect that the rigidity of singular terms can be expressed and thereby eliminated if we have quantifiers (and identity) at our disposal. Whether, in addition, something like criteria of cross-identification is needed or presupposed seems to me a secondary matter relative to this concern, although, I have to say, thinking about the role played in Kripke's account by individual essences, inclines

me to believe we were after something here. In any case, as I hope to make it clear below, I now think criteria of cross-identification are a secondary matter to questions of reference.

5. Rigidity, scope, and modal embedding

The fact that rigidity in the sense of constancy of designation can be expressed in formal languages with the help of quantifiers does not mean it is the correct way to capture the notion of rigidity for certain singular terms in natural languages. And it is this view which is in focus in *Naming and Necessity*. For those languages the mechanism consisting of quantifiers, variables, and binding, all in all, the "method of the variable," simply does not exist. Thus, it appears that the conclusion we drew, namely that it leads to the "dispensability of rigid designators or anything remotely like them in natural or formal languages" is not fully supported by the arguments we presented as they stand.

To be more precise, as I see it, there are two ways to counter our conclusion in the paper. Firstly, there is the claim that the expressibility of rigidity of singular terms with the help of quantifiers does not work for natural languages for the reason I just mentioned in the preceding paragraph. Secondly, there is the further claim that even if we had available scope distinctions, the expressibility strategy would not work simply because rigidity in natural language does not reduce to them, that is, is not a matter of scope distinctions.

I think that the first point can be easily taken care of: the scope mechanism can also be applied, although in a different format, to natural language to enforce rigidity. For instance, Dummett held the view that natural language has a convention according to which a name, in the context of any sentence, should be read with a large scope including all modal operators. The same idea, although in a different form and not applied to rigidity, appears in Hintikka's earlier work on the game-theoretical semantics (GTS) for natural language (Hintikka and Kulas 1985; Hintikka and Sandu 1991). GTS associates with a fragment of discourse a semantical game played by two players, Myself and Nature. Quantifiers, more generally logical expressions, and names prompt moves by

one of the players. A proper name prompts a move by Myself who chooses the referent of the name from the universe of discourse. Modal and intensional concepts are handled by combining game-theoretical semantics with possible worlds semantics. To take an example, the rule (G. knows that) looks like this:

- If the game has reached a sentence of the form "b knows that X" and a world w_1, then Nature may choose an epistemic b-alternative w_2 to w_1. The game is then continued with respect to X and w_2.

A specificity of natural languages, due to the lack of scope indicators, is that game rules must be complemented by a set of ordering principles which govern their order of application (cf. Hintikka and Kulas 1985, section 8). More importantly for the present purpose, the game rule for names, (G.name), has priority over many other game rules applicable to the constituents of the same clause. This amounts, in the traditional jargon, to proper names having "broader scope" over many other expressions in the same clause. True enough, the issue of the priority of proper names over the game rule for intensional operators has never been, to the best of my knowledge, systematically addressed in the GTS literature. My point in bringing it up is only to show that GTS has the resources to handle it. This way of handling it also shows, incidentally, that GTS assumes a convention about natural language according to which names have larger scope than many logical expressions and operators, including modal ones.

The "larger scope" view of rigidity in natural language has, however, been dismissed by Kripke, as somehow incoherent. As I mentioned in section 3, this view, held, among others, by both by Dummett and Hintikka, eliminates rigidity only in sentences with modal operators. In this connection, Kripke observes against Dummett that rigidity appears and makes sense not only in sentences with modal operators but also in simple sentences like "Aristotle was fond of dogs" (cf. our discussion in section 2). In other words, rigidity is a doctrine about the truth conditions of all kinds of sentences, simple and modal ones. Kripke acknowledges that the thesis of the rigidity of names in simple sentences can be expressed as

a "wide scope" phenomenon, that is, he agrees that that view is equivalent (ignoring complications arising from the possible nonexistence of an object) to the thesis that if a modal operator governs a simple sentence containing a name, the two readings with large and small scopes are equivalent (Kripke 1980, 12, fn. 15). But this equivalence, Kripke continues, "goes against the doctrine that natural language has a convention according to which only large scope reading is allowed. In fact, the equivalence makes sense only for a language where both readings are admissible" (ibid.). To conclude, the strategy we followed to eliminate rigidity in Hintikka and Sandu (1995) works only for sentences with modal embeddings. Rigidity, however, is a thesis about all kinds of sentences, including simple ones.

Perhaps I should add, commenting on the conclusion, that I believe Hintikka has never reached a definitive opinion on these matters. For instance, in Hintikka (1996), he reconsiders the difference between *de dicto* and *de re* epistemic attitudes. But now, somehow surprisingly, he uses the distinction to argue for the need for "rigid designators" in the language:

> …we need two kinds of singular terms. We need terms which pick up the same individual in all possible worlds; and terms which designate different individuals in different possible worlds. Constants proper serve the former purpose; ordinary (improper) serve the latter. Our improper constants are obviously related closely to Russell's logical proper names and to Kripke's 'rigid designators'. (Hintikka 1996, 122.)

Before closing the section, let me point out that the existence of rigidity in simple sentences (recall Kripke's example "Aristotle was fond of dogs") which shows its priority over modal embeddings, does not show in my opinion that there is no connection between naming and necessity, as claimed, e.g., in Almog (1986). As I observed in section 2, following Kripke, the rigidity of "Aristotle" in "Aristotle was fond of dogs" is manifest in the way we understand the truth conditions of this sentence in *counterfactual* situations.

6. Rigidity, quantifiers, and reference

It follows from what we said in the previous section that even independently of modal and attitudinal embeddings, the reduction of rigid reference to the objectual interpretation of quantifiers and quantifier scope is off the target, given that even in extensional fragments, reference and rigidity do not have to do with quantifier treatment, for the same reason they do not have to do with modal operators either. And when I say this, I have in mind natural languages. That is, reference concerns simple locutions in "Nixon is blue" or "John loves Mary," whereas quantification is semantically and logically posterior to the treatment of such simple nouns and predicates. If this is so, then the question of how to read and deal with quantifers is logically independent from the question of the semantical and logical analysis of proper names, which is prior to it. In other words, we should be free to interpret a quantifier objectually or substitutionally or blown it away altogether with no variables, with no constraints imposed by the interpretation of simple nouns and predicates. That is, there should be reference without objectual quantifiers and independently of the quantifying in into alethic, belief, or knowledge embeddings. The definability of the rigid reference of singular terms is a model-theoretical notion (in the sense of constant designation across a class of possible worlds) which may have a role to play in formal languages. I believe we were right about it in the case of formal languages with modal operators. For natural languages though, the model-theoretical expressibility is out of place and does not show the dispensability or eliminability of rigidity for a class of expressions or other.

University of Helsinki

References

Almog, Joseph (1986). "Naming without Necessity." *The Journal of Philosophy*, Vol. 83, No. 4, 210–242.

Davidson, Donald, and Gilbert Harman (eds.) (1972). *Semantics of Natural Language*. Dordrecht: D. Reidel Publishing Company.

Hintikka, Jaakko (1969). "On the Logic of Perception." In J. Hintikka, *Models for modalities*, Dordrecht: D. Reidel.

Hintikka, Jaakko (1996). "World Lines and Their Role in Epistemic Logic." In P. I. Bystrov and V. N. Sadovsky (eds.), *Philosophical Logic and Logical Philosophy*- Dordrecht: Kluwer, 121–137.

Hintikka, Jaakko, and Jack Kulas (1985). *Anaphora and Definite Descriptions*. Dordrecht: D. Reidel Publishing Company.

Hintikka, Jaakko, and Gabriel Sandu (1991). *On the Methodology of Linguistics*. Oxford and Cambridge, MA: Basil Blackwell.

Hintikka, Jaakko, and Gabriel Sandu (1995). "The Fallacies of the New Theory of Reference." *Synthese*, Vol. 104, No. 2 (Aug., 1995), 245–283.

Kripke, Saul (1963). "Semantical Considerations on Modal Logic." *Acta Philosophica Fennica* 16, 83–94.

Kripke, Saul (1976). "Is There a Problem about Substitutional Quantification?" In G. Evans and John McDowell (eds.), *Truth and Meaning: Essays in Semantics*. Oxford: Clarendon Press, 325–419.

Kripke, Saul (1980). *Naming and Necessity*. Oxford: Basil Blackwell.

The Semantics of Natural Kind Terms: A Critical Reflection on Experimental and Theoretical Issues[1]

GENOVEVA MARTÍ

1. Introduction

The approach to natural kinds and to the semantics of natural kind terms defended by Kripke and Putnam (from now on, the KP approach) has been discussed, objected to, and defended, since it was proposed in the early 70's.[2] Kripke and Putnam endorse an externalist (or causal-historical) approach to semantics, according to which facts that are beyond the cognitive grasp of a competent speaker can contribute to the determination of the reference of the speaker's use of a term. Kripke's arguments and, in particular, Putnam's Twin Earth

[1] A disclaimer about the title is in order. There are interesting metaphysical issues as regards natural kinds. However, I do not think natural kind *terms* constitute a distinctive *semantic* category. "Tiger," "gold," "pencil," or "philosopher" in my view behave semantically the same way, namely, they designate kinds or attribute membership to kinds. I think that this applies also to so-called social kind terms. As I will argue below, this does not entail that there are no descriptive kind terms. In this paper I focus on a discussion that raises issues about some natural kinds, in particular, biological kinds, and also about the use of those terms, hence I will often fall in line with the tradition of talking about "the semantics of natural (or biological) kinds." I thank Katarzyna Kijania-Placek for discussion of this issue and Andrea Bianchi for prompting me to address it.

[2] See Kripke's 1970 lectures (1980), especially lecture 3, and Putnam 1973 and 1975. There are important differences between Kripke's and Putnam's respective stances, for instance as regards the role of the appeal to experts (see Kripke 1986 for discussion). Those differences will not be relevant for this paper, but it should be kept in mind that talking about "the KP model" or "the KP approach" is an oversimplification.

story, are meant to dislodge the classical descriptivist paradigm that was generally accepted at the time especially as regards terms such as "gold," "water" or "tiger" (as an aside, it should be noted that for Putnam, at least at some point in time, the considerations against descriptivism applied also to terms such as "pencil").

The debate around the KP approach has taken different forms. Some authors have argued for or against the model putting forward arguments that focus on the scientific practices of naming and classifying in different disciplines, or on the theoretical commitments of specific scientific theories. For instance, one of the early dissenters, John Dupré (1981), examining how classification is conducted in the biological sciences, argued that the KP model was inadequate, and others have brought to the fore arguments that rely on scientific practice in chemistry, physics, and other disciplines.[3]

The debate has been conducted also on the basis of experiments that seek to collect data by asking participants in the experiment to respond to questions after being exposed to stories similar to the ones envisaged by Putnam in the Twin Earth scenario. The discussion based on this methodology is not just a recent phenomenon circumscribed to philosophers. Although some studies on categorization led entirely by psychologists obtained results that suited some aspects of the KP approach (see for instance Rips 1989), other studies (see for instance Braisby, Franks and Hampton 1996) obtained results that were not in line with what are taken to be crucial assumptions of the model.[4]

The discussion of the KP model has taken another turn as of recent with the publication of some studies by experimental philosophers (some of them conducted in collaboration with psychologists) on biological kind terms. Following the strategy exemplified by Braisby and colleagues and other psychologists, experimental philosophers test the general population by presenting them with stories involving natural kinds and deriving from their responses some conclusions

[3] The list of disputants for and against is extremely long.

[4] Braisby, Franks, and Hampton focus on the role of essence in categorization. Doubts as to whether essentialism is a fundamental commitment of KP's semantic model are discussed below.

about the use people make of the terms that designate those kinds. This is in line with the methodology applied by experimental philosophers, in general, to test whether the counterfactual scenarios that philosophers envisage to reach what they take to be intuitive conclusions (for instance about the correct application of a term) provoke the same kind of reaction among the population at large.

In this paper I will discuss critically some of the conclusions presented in recent studies performed by experimental semanticists.[5] Before focusing on the discussion, and in order to put some issues in context, I present some general reflections about experimental philosophy in general.

2. Some remarks about experimental philosophy and "the armchair"

The vast majority of experimental philosophy studies consist in telling people a story (or having them read a vignette) and then asking them certain questions. Experimental philosophers often describe their objectives as "testing the intuitions" of a population, to determine whether their intuitions and those of professional philosophers coincide, or to test whether experts agree in what they consider an intuitive response. This has led to interesting discussions of what intuitions are or what kinds of intuitions are relevant.[6] I will not engage in that discussion because it seems to me that the kinds of tests in which people are given a vignette and then answer some questions test what I would characterize as *initial reaction*s or *initial responses* to the story told. In those tests, participants are presented with a story and then they are expected to provide the answers that seem natural to them. So, I believe it is right to think of the data collected as initial responses. And I am using "initial" because the declared objective of experi-

[5] In the past I have argued in support of the KP model for natural kinds and for kind terms. See Hoefer and Martí 2020 and Hoefer and Martí 2019 which is a response to Häggqvist and Wikforss 2018. Other participants in this very recent debate include Raatikainen 2021 and Häggqvist 2022. None of these discussions involve arguments about experimental philosophy tests.

[6] A lot of the debate is inspired by Williamson 2004 and Devitt 2010.

mental philosophers is, in fact, not to collect heavily reflected on data.

So, this raises an issue: what should we philosophers do with those initial reactions? This is a question that deserves some thought, because we often find experimental philosophers claiming that the results of their tests should have consequences for philosophical theories. Just to give a couple of examples: Machery, Mallon, Nichols and Stich in their seminal 2006 article claim that their results raise "questions about the nature of the philosophical enterprise of developing a theory of reference" (B1). And Cova et al. (2019), after performing tests on aesthetic judgements, conclude that "the traditional way of approaching the debate over the nature of aesthetic judgement is fundamentally misguided" (Cova et al. 2019, 335) and that "philosophical inquiries about the nature of aesthetic judgments should no longer take [certain assumptions] as a starting point" (ibid., 337).

I often teach philosophy of language and I explain to students that an essential part of semantics is the theory of truth conditions. And since it is important to get clear about what we mean by "truth conditions" I ask the students this question: "If we called birds 'pigs', would pigs fly?"[7]

About 80% of the students raise their hand: yes, if we called birds "pigs," pigs would fly. And the majority of the remaining 20%, I suspect, don't react because this must be a tricky question and "the obvious answer" may not be right. So, what do I do with this? What do I do with their initial reaction? Well, I discuss it and I reflect with them.

I do not conclude that evidence collected year after year of teaching introductory philosophy of language supports the claim that laypeople think that all you need to do to make a pig fly is just a matter of changing the words we use.

I proceed to explain that when we ask ourselves whether what we say when we use a given sentence would be true under different circumstances (i.e., whether what I say when I utter "pigs fly" would be true in the circumstances described) we are not asking whether the sentence, if uttered under different circumstances would be true, or express a truth. We are

[7] That, by the way, was one of the questions asked to applicants to the undergraduate degree of Philosophy at Oxford University.

asking whether what we in fact say would be true in a scenario that differs from actual circumstances only in the fact that birds are called "pigs."

It doesn't take too long for my students to see that they interpreted the question as the question whether "pigs fly" would express a truth if uttered in a scenario in which we called birds "pigs," and that this interpretation is not what we are after when we ask ourselves about the truth conditions of our utterances of "pigs fly."

When they understand that, they understand what it means to say that if two utterances of sentences have different truth conditions, they must be expressing different things, and they can thus master tools that we need to advance in our philosophy of language course.

And of course, they also learn that the only way pigs could fly would be for them to grow wings (something that, I don't doubt, they knew all along).

All this suggests, in my view, that it is not even clear at all that people's initial reactions are evidence of what they really think. As philosophers we need to ask ourselves what we can use as the raw material to start the philosophical enterprise: immediate, knee-jerk reactions, or subsequent reflective responses?

In any case, although knowing the initial reactions of my students is extremely useful (among other things, it alerts me of confusions that need to be resolved), the data does not have, and should not have, an impact on the theory of truth conditions. Philosophical, and philosophically guided, reflection on the data is necessary. In general, rather than attempting to base or debunk philosophical theories by appeal to the kind of data collected in experimental philosophy surveys of initial reactions, it might be more fruitful to think about the data in question as the starting point to deliberate on the sorts of considerations that once highlighted lead to reflective and reasoned responses on the part of the participants in experimental tests, and a fortiori, on the part of the general population.[8]

---

[8] That is not just an abstract philosophical point. I believe that experimental philosophers have the responsibility to clarify their stance on this issue, especially in an era of instant, non-reflected, evidence-blind opin-

It is tempting to conclude that, although experimental philosophy may provide interesting data for philosophical reflection, so-called *armchair philosophy* continues to have a decisive role. I myself would be happy with that conclusion if it weren't because I am not sure what the term "armchair philosophy" is supposed to apply to. The papers mentioned in the previous section, pieces that engage in a debate on metaphysical and semantic issues involving kinds and kind terms, present arguments based on scientific theories and consider examples taken from past and recent history of science. Putnam (1975) himself has physical and chemical facts and theories very present in his arguments. And in 1990, justifying the simplification of regarding water as essentially constituted of molecules of H_2O he writes: "I shall stick to high school chemistry because the actual quantum-mechanical picture of the structure of water is immensely complicated" (ibid., 57, fn. 3). These works present philosophical reflections on scientific results and on scientific theories. I am not sure if experimental philosophers regard them as products of armchair theorizing. And if they do, why that is so. In any case, the "armchair philosophy" metaphor needs sharpening.[9]

3. Biological kind terms. Experimental and theoretical issues

There have been as of late several experimental studies on the use of kind terms, often with widely different results. Some of those studies report substantial disagreement among partici-

---

ions and reactions. Of course, this is not to say that knowing the immediate, unreflective responses of people are never of value to philosophical reflection (see footnote 15 below).

[9] A related issue is raised by Brian J. Scholl (2007) who expresses the concern that the traditional experimental philosophy studies that consist in having participants read vignettes and answer questions "rather than telling us anything about underlying mental mechanisms, may instead often tell us more about how subjects respond to bizarre questions and scenarios." And he encourages instead experiments that use "more implicit response measures that help to ensure that the results reflect underlying mental mechanisms…" (580–581). For a discussion of the effects of the failure to distinguish implicit mechanisms from explicit responses in a particular study see Contesi et al. (forthcoming).

pants and even a good number of contradictory responses by individual participants. In this paper I will discuss the conclusions of some these studies and reflect on their impact on the theory of reference for kind terms. The literature on this topic is rather extensive, so I will focus on some of the most recently reported results on the use of biological kind terms.

Haukioja, Nyquist and Jylkkä (2021) as well as Devitt and Porter (2021) use a mixture of elicited production or EP (where people are asked to use the terms being studied) and truth-value judgements or TVJ (where participants are asked to answer "true" or "false" when prompted with some sentences). Although Devitt and Porter ultimately criticize some aspects of the methodology followed by Haukioja, Nyquist and Jylkkä, both studies agree in concluding that "both mainstream externalist and traditional internalist theories of reference are mistaken" (Haukioja, Nyquist and Jylkkä 2021, 401) and so that "we should abandon the common assumption that any one theory of reference fits all natural kind terms" (Devitt and Porter 2021, 1) because "there are indeed *both* descriptive and causal historical elements to the reference determination of biological kind terms" (Devitt and Porter 2021, 27). A more recent article draws similar conclusions from further tests (Devitt and Porter, 2023).

In a prior study, involving proper names, Michael Devitt and Nicolas Porot (2018) had used elicited production and truth value judgments. The use of elicited production was particularly important since their study came in the heels of prior surveys that obtained results in line with the predictions of a descriptivist approach to the semantics of names, but that relied heavily on questions eliciting referential judgements from participants, i.e., questions that constituted evidence of the participants' opinions as regards what uses of names referred to, not evidence of how they themselves used the names. Performing tests that did target the participants' usage of proper names Devitt and Porot obtained results substantially consistent with the causal-historical non-descriptivist picture.

In extending the Devitt and Porot methodology from singular to kind terms, Devitt and Porter tell us that their hope was that the correct methodology would confirm the results that Devitt and Porot had obtained using similar methods for

proper names, results that gave overwhelming support to the causal-historical picture.

But the results of the tests with biological kind terms came as a surprise: "The results… were neither what we expected nor what we had hoped for. Far from showing that the Kripke–Putnam causal-historical theory is correct after all, they confirmed the main conclusions of earlier… tests: Reference is to be explained partly descriptively and partly causal-historically (nondescriptively)" (Devitt and Porter 2021, 9).

In their 2021 paper Devitt and Porter perform an EP test in which, after presenting participants with a vignette, they put forward two statements, one of which corresponds to a descriptivist take on the story and another one that corresponds with a non-descriptivist take. And they also perform two TVJ tests in which each group of participants is given one statement, descriptivist or anti-descriptivist and asked whether the statement is true or false.

On the basis of the results, Devitt and Porter examine different proposals as to how the reference of biological kind terms is to be accounted for: an ambiguity theory or a hybrid theory and they ultimately defend a hybrid theory. I will not discuss these proposals to focus exclusively on the test and the surprising results.

Thus, consider some of the results of some of the tests performed by Devitt and Porter:

> 1. Faced with both nondescriptivist and descriptivist options at once, participants' choices were *close to 50–50*, with only an insignificant preference for the nondescriptivist one […]
>
> 2. Faced with the nondescriptivist statement without having been presented with the descriptivist statement, an extremely significant proportion of participants chose the *nondescriptivist* one […]
>
> 3. Yet, faced with [the] descriptivist statement without having been presented with the nondescriptivist statement, a highly significant proportion of participants chose the *descriptivist* one […] (Devitt and Porter 2021, 17)

These results, they claim, support strongly the presence of "descriptivist and non-descriptivist reference determination

of biological kind terms" both within the community and also within individuals (ibid., 17).

Some confusions about the theoretical assumptions underlying the discussion of natural kind terms, independent of the experimental issues raised in these papers, are worth mentioning and should be avoided.

The disagreement between descriptivist, or internalist, and causal-historical anti-descriptivist, or externalist, approaches to semantics is presented by Devitt and Porter (2021) as follows:

> [According to causal-historical theories, a] biological term like "tiger" does not refer to an animal in virtue of its having the superficial properties picked out by speakers' associated descriptions but rather in virtue of its having the same deep structural properties (the same underlying "essence")… (ibid., 2).

It is common to associate the causal-historical picture to the postulation of deep natures or essences. Devitt and Porter (2021) also endorse the association, and so do Haukioja et. al. (2021). The latter often mention in their discussion "evidence of ambiguity between superficial and deep features in categorization" (ibid., 396) as a sign of the internalist and externalist pull in different directions. But this is based on a confusion, on two counts.[10]

First, the description associated with a term may well be a description of the deep nature of a kind or a substance. Nigel Sabbarton-Leary (2010) mentions the case of the term "tungsten." The meaning of "tungsten" is given by the description that captures the essence of tungsten: "the element with atomic number 74." Any application of the term "tungsten" to a sample that does not satisfy the description is just incorrect and incompetent. So, obviously a descriptivist approach to reference is not contrary to the postulation of deep natures, and it does not automatically deny them any role in the determination of reference.

Second, we should not forget that a crucial component of Putnam's approach is the idea that we classify by similarities.

[10] The confusion affects not only the debate in experimental philosophy; it is pervasive and so, it is worth clarifying it.

And the similarities in question may well not be deep structure, although the appeal to deep structure is a way to argue for the externalist stance that *meaning ain't in the head* (or at least not all of it).

Martí and Ramírez-Ludeña (2016) put the point as follows:

> It is often taken for granted … that the Kripke–Putnam approach to the semantics of general terms is committed to essentialism, the postulation of shared underlying natures that are not immediately accessible or observable and can be discovered only by scientific investigation. But the commitment to essentialism is not constitutive of the approach. On the Kripke–Putnam model some samples or individuals are treated as paradigms, and other instances are classified as members of the same kind by virtue of their similarity to the paradigms. The similarity could well be superficial (based on how new yet to be classified objects or samples appear or look), or based on sameness of function. The Kripke–Putnam model does not impose that the relevant criterion is essence. The novelty of the view is rather that it *opens the door* to the possibility that the similarity that is responsible for certain classifications into kinds be entirely external to the minds of speakers. (Martí and Ramírez-Ludeña 2016, 126)

And of course, the appeal to the microstructure of water in the Twin-Earth case makes the point dramatically, since hardly anything could be more out of cognitive access than a yet unknown microstructure.

In any case, the dissociation of the externalist stance from the postulation of the role of shared underlying natures is not just a charitable re-interpretation. Putnam himself was very clear on this:

> Another misunderstanding that should be avoided is the following: to take the account we have developed as implying that the members of the extension of a natural-kind word necessarily *have* a common hidden structure. It could have turned out that the bits of liquid we call "water" had *no* important common physical characteristics *except* the superficial ones. In that case the necessary and sufficient condition for being "water" would have been possession of sufficiently many of the superficial characteristics. (Putnam 1975, 159)

In the recent article, Devitt and Porter (2023, 6) report that Andrea Bianchi has alerted them in conversation of the inaccuracy of the association between the causal-historical approach and the commitment to reference being fixed by deep structural properties. Devitt and Porter report that the issue does not affect their results, since they suggest descriptivist and anti-descriptivist leanings on the part of participants, even without the assumption of underlying natures (ibid., 18). They do not report if they have also taken into account the dissociation of descriptivism and superficial features mentioned here. Namely, they do not report if descriptivist and anti-descriptivist leanings on the part of the participants are detected *on* the assumption of underlying natures. In any case, independently of whether the results of the Devitt and Porter experiments can be considered robust, the theoretical point stands: the quick association of the externalist stance and the appeal to hidden essence is, indeed, too quick.[11]

In any case, Devitt and Porter (2021 and 2023) and Haukioja, Nyquist and Jylkkä (2021) claim that people are pulled in different directions: the causal-historical direction when then they classify samples according to their deep nature, and the descriptivist direction when they classify according to superficial features.

There are some hypotheses about why and when this happens. Tobia, Newman and Knobe (2020) suggest that the variation is driven by context. Participants that had to judge whether something was a salmon tended to rely on superficial features in legal scenarios, something that appears to suggest that in practical contexts uses of kind terms are con-

[11] To be more precise, we should also distinguish the distinction deep/superficial from the distinction essential/accidental. There is nothing in principle wrong with a view according to which some essential properties are superficial and observable. On the other hand, the claim that microstructural properties, such as having the molecular structure H_2O are important physical properties that classify certain samples as samples of water, does not by itself automatically entail that the property in question is a necessary property of the kind (nor of the sample, obviously, but that is beyond doubt). Plausible as the association deep/essential might be, a subsequent metaphysical argument is required. In general, the discussion surrounding the KP model takes for granted the association without finer distinctions.

sistent with the predictions of descriptivism, but Devitt and Porter (2023) find no evidence supporting that hypothesis: according to their results the variations are not driven by context, but rather by whether a term is or is not of practical interest. In their tests Devitt and Porter (2023) compare a term with no practical interest ("Rio de Janeiro Myrtle") with a common term with obvious practical interest ("rice") and they report that whether in scientific or practical scenarios, "the results support a Causal-Historical Theory of 'Rio de Janeiro Myrtle' and are evidence against a Causal-Historical theory of 'rice'" (ibid., 18).[12]

It is not my purpose here to discuss the details and relative merits of the different studies. But one aspect of the "rice" case invites reflection.

One of Devitt and Porter's vignettes tells the story of a synthetically created seed, that has the same look, taste and nutritional content[13] as *Oryza sativa* (rice) but a completely different genetic structure. A lab assistant takes a bag of the new seed to a restaurant where the chef serves it as rice. And the question is whether what the chef serves is rice.

Although the responses are significantly more in accord with the causal-historical approach, there is a substantial minority of descriptivist answers, supporting the general conclusion, according to Devitt and Porter, that there are both causal-historical and descriptivist elements in the determination of the reference of "rice."

Devitt and Porter, in their 2021 paper are surprised at the proportion of uses that seem to be guided by a definite description associated with the terms tested.

But, how much a surprise should that be? I don't think it should be surprising to us that people be ready to put together things according to the features that are important to them, in particular if the term in question is what Devitt and Porter qualify as a term "of practical interest," and often superficial

[12] It is hard to tell if these results will be confirmed further. Cases such as the different uses of "fruit" established in the community (culinary and botanical) seem to be clearly contextual.

[13] Nutritional content is not a superficial feature, but it is certainly a feature known by the general population and hence, cognitively accessible.

features are important. They are the features that we use every day to identify things.

It is not clear either that Putnam himself would be surprised. In Putnam 1975 we read: "… in one context 'water' may mean *chemically pure water*, while in another it may mean the stuff in Lake Michigan. And structure may sometimes be unimportant; thus one may sometimes refer to XYZ as water if one is *using* it as water" or "we discover 'tigers' on Mars. That is, they look just like tigers, but they have a silicon-based chemistry rather than a carbon-based chemistry… Are Martian 'tigers' tigers? It depends on the context". (Putnam 1975, 157–158).

Now, Putnam seems to assume that the variation in usage depends on context. Devitt and Porter conclude from their experiments that the variation in question is not driven by context, and thus they defend a type of hybrid approach to the semantics of natural kind terms, one that incorporates features of the causal-historical picture and features of descriptivism, features that, sometimes and for different people (and even for the same person), pull in different directions. Their results put pressure on the context-driven explanation of the variability proposed by Tobia, Newman and Knobe (2020). As I said, I will not discuss here this aspect of the debate.[14] The point is that the variability in the use of kind

[14] Devitt and Porter recruited their participants through MTurk, and the results of their test indicate that those participants used "rice" in ways that accord with the causal-historical view and in ways that accord with descriptivism both in a practical context (the one involving a restaurant that serves the new seeds as rice) and in a scientific context (one in which the seeds are taken to a botany class as rice seeds). In fact, there were more responses in line with descriptivism in the scientific context. The presence of descriptivist responses in both contexts is the basis of Devitt and Porter's argument against a contextually driven approach and in favor of a hybrid approach. Perhaps it would have been good to know, though, how botanists themselves would use "rice" in each context. This is, of course, anecdotal evidence, but I think that even expert botanists understand that when we ask them if they put fruit in their salads (a "practical" context), we are asking them if they put apples, pears, strawberries, etc., and we are not asking them if they put tomatoes. But I doubt that in a "scientific" context any of them would argue that tomatoes do not belong to the botanic category of fruits.

terms, sometimes driven by appearance and accidental features, and sometimes driven by the assumption of a common nature, is not a surprise, not even for Putnam.

The observation of the variation in usage leads to the conclusion that "both mainstream externalist and traditional internalist theories of reference are mistaken" (Haukioja, Nyquist, and Jylkkä 2021, 401) and "we should abandon the common assumption that any one theory of reference fits all natural kind terms" (Devitt and Porter 2021, 1).

The presumption here is that the externalist causal-historical position denies that there can be uses of kind terms governed by cognitively accessible definite descriptions. Why else would the presence of responses consistent with descriptivism suggest that externalism is mistaken?

As I have argued in the past, this is to misunderstand the dialectic between descriptivism and anti-descriptivism (Martí 2015, 2020): "Descriptivism is a hegemonic approach to reference. It postulates that reference is always mediated by a definite description: it is *impossible* to refer without the mediation of descriptive material, cognitively accessible to the speaker, that determines the reference, or domain of application, on each occasion of use" (Martí 2020, 337).

The externalist arguments used by Kripke, Putnam and others show that it is *possible* to refer without the mediation of a cognitively accessible definite description, that, as Keith Donnellan (1970) put it, a backup of descriptions is neither necessary nor sufficient to refer. The arguments are not supposed to show that terms cannot refer, ever, via associated descriptions. Results that show that some uses are guided by the descriptive material people associate with a term are interesting as a report of how people use language, and as such they invite a philosophical reflection. It may be that for some terms, or some classes of terms, application is semantically guided by definite descriptions. But that does not mean that neither internalism/descriptivism nor externalism/anti-descriptivism are entirely correct as Devitt and Porter or Haukioja, Nyquist and Jylkkä's conclude. The externalist (unlike the descriptivist) never assumed that all terms have to fit one mold.

It should be observed also that Putnam didn't take back his Twin Earth case when he acknowledged that we might de-

cide to call XYZ "water." This is, I contend, because deep down, the important point was always metaphysical: whether a substance whose molecular composition was largely XYZ was the same substance as water or a different kind of thing. It is undeniable that both Kripke and Putnam present their views couching the fundamental points in terms of language and meaning. This may be a reflection of the status that language had originally in analytic philosophy as the key to metaphysical, epistemological and ethical issues. I think though that the underlying fundamentality of the metaphysical point is revealed by the fact that Putnam does not revise the Twin-Earth case when he contemplates other uses of "water."

As regards the "rice" case, one wonders what would happen if the story presented to the participants was: you are in a restaurant where the new seeds (those that a substantial portion of people in Devitt and Porter's experiment have no doubt in classifying as rice or a new type of rice) are served in dishes that, in the menu, appear as containing rice, and you are having lunch with your very good friend who is severely allergic to most foods. But she can safely eat rice. Will you order "rice" for both of you?

Similarly, suppose that we call XYZ "water." After all, we shower with it, we wash dishes with it, we even ingest it orally. But suppose we have never had it injected directly into our veins and that no research has been done before to test how the XYZ molecule interacts when human blood is exposed directly to it. If you are severely dehydrated, will you happily acquiesce to having an XYZ saline solution drip?

I, for one, wouldn't recommend the dishes that, according to the menu, contain rice to my friend nor would I happily accept the XYZ drip, without further research using that seed or that substance.

Amie Thomasson (2020) puts the point in terms of concepts, but the claim travels easily to the categorization of kinds:

> I have a child with a nut allergy. It is a matter of life and death ("death in seven minutes", her allergist tells us) whether something is biologically a tree nut or is something *called* a 'nut'. It is a matter of life and death because it enables us to *predict* whether ingesting something will cause a life-threatening allergic reac-

> tion. It is not just a subjective matter whether 'tree nut' is a better concept than one that includes all and only things called 'nut' (including hazelnuts, peanuts, coconuts, nutmeg, and doughnuts (only the first of which is biologically a tree nut), and excluding cashews, pistachios, and almonds). That one concept but not the other is usefully and efficiently *predictive* in this way, which has life-or-death consequences, is all I need to be fully convinced that one set of concepts is objectively better. (Thomasson 2020, 450)

In general, the stories used in experimental philosophy tests, Devitt and Porter's in particular, do not describe high-stakes, life and death scenarios in which decisions have important consequences, consequences that involve us or someone very close to us. So the scenarios do not invite serious reflection. Participants in the experiments are not invited to think hard. They give unreflective responses, in part because the explicit aim of these studies is to collect immediate reactions. And the direct value of immediate, unreflective, reactions to philosophical theorizing is often questionable.[15]

Now, I do not know if these considerations speak in favor of further tests in which life and death stories are presented. I only know that if the results of potential new tests that take these issues into account contradict me, if it turns out that people would happily accept an XYZ drip or would gladly recommend their seriously allergic friend to have "rice," I myself would not change what I think right now. In the circumstances envisaged, I would not have XYZ injected, and I would not risk hurting my friend. For the point is that, without further scientific testing, we would not know if XYZ, or

[15] This is not to say that immediate reactions are never useful as input for philosophical reflection. For certain purposes, they may be exactly what is required. For instance the psychological tests on generics by Cimpian, Brandone and Gelman (2010) elicit immediate reactions that show that people judge that the proportion of satisfaction of a property attributed to members of a group by a generic statement is very high, while at the same time they are ready to judge a generic true on the basis of a much lower amount of satisfaction of that property by members of the group. These data certainly invite a philosophical, and social, reflection on the acceptance of generics about human groups. See also Cella, Marchak, Bianchi and Gelman 2022 for discussion.

the new "rice" have some, so far, unobserved harmful effects. Two sorts of stuff having different underlying constitution may have displayed all the same observable behavior so far. But, in general, we cannot expect that the same behavior will continue in all future contexts.[16] And that's because they are different kinds of things, whether we call them with the same name or not. So, it makes a lot of sense to be cautious.

And, I think, at least some of the participants in Devitt and Porter's studies are in fact quite conscious of that. Devitt and Porter report that they had to modify the vignette because several participants thought that "we were asking them to judge the morality or legality of the chef's actions" (Devitt and Porter 2023, 10). Why would it be immoral or illegal for the chef to serve "rice," if people's use of the word "rice" had always included that seed, as the descriptivist leanings Devitt and Porter detect in the population appear to suggest?

Devitt and Porter modified the restaurant vignette adding "Leaving aside whether this is an appropriate thing for the chef to do… "It is interesting that the actions of the lab assistant that takes a bag of the new seeds from the lab without asking for permission is not a matter of concern in the restaurant vignette, nor in another vignette in which the lab assistant takes the new seeds to a botanics class, also without asking for permission. The concerns are raised exclusively as regards the actions of the chef that serves the new seeds in dishes that, according to the menu, contain rice.

If we have two kinds, it is usually wise to have two words. Of course, this is not always the case. We use "jade" for two different minerals. Nephrite and jadeite have a fundamentally ornamental value, so it may not be important to use different words for them in everyday life. Would we accept to use "rice" for the new seeds if there was the possibility that its different genetic structure provoked unexpected side effects (something that the Devitt and Porter vignettes never bring up)? Would we, if sufficient research definitely showed that the new seeds were as harmless as rice? Perhaps. Paraphrasing Putnam, if tastes as rice, looks like rice and we use it as rice, we may call it "rice." And we might call XYZ "water"

[16] See Hoefer and Martí 2019, section 5 for a discussion of this issue in relation to the Twin Earth case.

and Martian tiger-look-alikes "tigers." After all, we call two different minerals "jade."[17]

But the use of one word or two words should not mask the fundamental issue that rice and the new synthetic "rice" are different kinds of things. Animals with different biological histories, minerals with different compositions, substances with different molecular microstructures and seeds with entirely different origins are different kinds of things. And the predisposition of people to use one word for two kinds (a predisposition that, in my view, has not been properly tested by Devitt and Porter because of their reliance on unreflective responses to humdrum stories) can do nothing to alter the more fundamental fact.[18].

ICREA and Universitat de Barcelona

References

Braisby, Nick, Bradley Franks, and James Hampton (1996). "Essentialism, Word Use and Concepts." *Cognition* 59, 247–274.

Cella, Federico, Kristan A. Marchak, Claudia Bianchi, and Susan A. Gelman (2022). "Generic Language for Social and Animal Kinds: An Examination of the Asymmetry between Acceptance and Inferences." *Cognitive Science* 46(12), e13209.

[17] What to say about the envisaged uses of "water," "tiger" or "rice"? Some may be tempted to chalk them off as pragmatics. But the point is that if the uses become established in the community it is not clear at all that the pragmatic explanation is satisfactory. Take, for instance, the case of "fruit." Even the courts have acknowledged that there are two uses of "fruit" established in the community. One of them, botanical, applies to tomatoes. The other one, culinary, does not. This is not to be dismissed as a purely pragmatic phenomenon. Is the case of "fruit" a case of ambiguity? Is a hybrid explanation required? Is it rather a case of polysemy? For the case of "fruit" at least, ambiguity would seem to be a plausible answer, given that the two uses of "fruit" are lexicalized differently in other languages (in Greek, for instance).

[18] Versions of this paper were read at the Warsaw Sign-Language-Reality Seminar and at the Jagiellonian University (Krakow). I am grateful to the audiences for their comments.

Cimpian, Andrei, Amanda Brandone, and Susan A. Gelman (2010). "Generic Statements Require Little Evidence for Acceptance but Have Powerful Implications." *Cognitive Science*
34(8), 1452–1482.

Contesi, Filippo, Enrico Terrone, Marta Campdelacreu, Ramón García-Moya and Genoveva Martí (forthcoming). "The Paradox of Taste to the Experimental Test." *Analysis*.

Cova et al. (2019). "De Pulchritudine Non Est Disputandum? A Cross-cultural Investigation of the Alleged Intersubjective Validity of Aesthetic Judgment." *Mind and Language* 34, 317–338.

Devitt, Michael (2010). "Linguistic Intuitions Revisited." *British Journal for the Philosophy of Science* 61, 833–865.

Devitt, Michael, and Nicolas Porot (2018). "The Reference of Proper Names: Testing Usage and Intuitions." *Cognitive Science* 42(5), 1552–1585.

Devitt, Michael, and Brian C. Porter (2021). "Testing the Reference of Natural Kind Terms." *Cognitive Science* 45(5), e12979.

Devitt, Michael, and Brian C. Porter (2023). "Two Sorts of Biological Kind Terms: The Cases of 'Rice' and 'Rio de Janeiro Myrtle'." *Philosophy and Phenomenological Research*. doi: 10.1111/phpr.12979

Donnellan, Keith (1970). "Proper Names and Identifying Descriptions." *Synthese*, 21, 335–358.

Gómez-Torrente, Mario (2019). *Roads to Reference. An Essay on Reference Fixing in Natural Language*. Oxford: Oxford University Press.

Häggqvist, Sören (2022). "No, Water (Still) Does Not Have a Microstructural Essence." *European Journal for Philosophy of Science*, 12(2), 1–13.

Häggqvist, Sören, and Åsa Wikforss (2018). "Natural Kinds and Natural Kind Terms: Myth and Reality." *British Journal for the Philosophy of Science* 69(4), 911–933.

Haukioja, Jussi, Mons Nyquist, and Jussi Jylkkä (2021). "Reports from Twin Earth: Both deep structure and appearance determine the reference of natural kind terms." *Mind and Language* 36, 377–403.

Hoefer, Carl, and Genoveva Martí (2019). "Water Has a Microstructural Essence After All." *European Journal for Philosophy of Science* 9, 12.

Hoefer, Carl, and Genoveva Martí (2020). "Realism, Reference and Perspective." *European Journal for Philosophy of Science*, 10(3), 1–22.

Kripke, Saul A. (1970/1980). *Naming and Necessity.* Cambridge, MA: Harvard University Press.

Kripke, Saul A. (1986). "A Problem in the Theory of Reference: The Linguistic Division of Labor and the Social Character of Naming." In *Philosophy and Culture, Proceedings of the XVIIth World Congress of*

Philosophy. Montreal: Editions du Beffroi, Editions Montmorency, 241–247.

Machery, Edouard, Ron Mallon, Shaun Nichols, and Stephen Stich (2006). "Semantics Cross-cultural Style." *Cognition* 92, B1–B12.

Martí, Genoveva (2015). "General Terms, Hybrid Theories, and Ambiguity: A Discussion of Some Experimental Results." In J. Haukioja and J. Beebee (eds.), *Advances in Experimental Philosophy of Language.* London & New York: Bloomsbury, 157–172.

Martí, Genoveva (2020). "Experimental Semantics, Descriptivism and Anti-descriptivism. Should We Endorse Referential Pluralism?" In A. Bianchi (ed.), *Language and Reality from a Naturalistic Perspective*. Cham: Springer, 329–341.

Martí, Genoveva, and Lorena Ramírez-Ludeña (2016). "Legal Disagreements and Theories of Reference." In Alessandro Capone and Francesca Poggi (eds.), *Pragmatics and Law. Philosophical Perspectives*. Cham: Springer, 121–139.

Putnam, Hilary (1973). "Meaning and Reference." *The Journal of Philosophy* 70(19), 699–711.

Putnam, Hilary (1975). "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7, 131–193.

Putnam, Hilary (1990). "Is Water Necessarily H_2O?" In J. Conant (ed.), *Realism with a Human Face*. Cambridge, MA & London: Harvard University Press, 54–79.

Raatikainen, Panu (2021). "Natural Kind Terms Again." *European Journal for Philosophy of Science* 11(1), 1–17.

Rips, Lance J. (1989). "Similarity, Typicality and Categorisation." In S. Vosniadou and A. Ortony (eds.), *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press, 21–59.

Sabbarton-Leary, Nigel (2010). "Descriptivist Reference from Metaphysical Essence." *Dialectica* 64(3), 419–433.

Scholl, Brian J. (2007). "Object Persistence in Philosophy and Psychology." *Mind & Language* 22, 563–591.

Thomasson, Amie (2020). "A Pragmatic Method for Normative Conceptual Work." In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, 435–458.

Tobia, Kevin, George Newman, and Joshua Knobe (2020). "Water Is and Is Not H_2O." *Mind and Language* 35, 183–208.

Williamson, Timothy (2004). "Philosophical 'Intuitions' and Scepticism about Judgment." *Dialectica* 58(1), 109–153.

Type Specimens and Reference[1]

MICHAEL DEVITT

1. Introduction

In an ingenious and provocative paper, "Individualism, Type Specimens, and the Scrutability of Species Membership", Alex Levine argues that "species membership, by which I mean the relation that connects a given organism, *o*, with the species *S* of which it is part, is a fundamentally contingent matter" (2001, 333). He finds this contingency in conflict with the role of "type specimens" in biology. He points out that "naming a species requires collecting and preserving one, or at most a very few specimens of the species in question" (327). David Hull has the following view of this practice:

> The sole function of the type specimen is to be the name bearer for its species. No matter in which species the type specimen is placed, its name goes with it. (Hull 1982, 484)

Levine takes Hull's view, together with the "rigid designation" theory of reference, to entail that any organism selected as the type specimen for a species is necessarily a member of that species. This generates the conflict that Levine sums up neatly as follows: "*qua organism*, the type specimen belongs to its respective species contingently, while *qua type specimen*, it belongs necessarily"; he finds this "paradoxical" (Levine 2001, 334).

What precisely is Levine's necessity thesis about type specimens? Joseph LaPorte (2003) has clarified this question. He starts with the following statement of the thesis: "It is necessary that any species with a type specimen contains its type specimen". He points out that such statements have two readings:

---

[1] A version of this paper appears as chapter 5 in Devitt 2023.

> The *de dicto* reading of the statement in question would typically be expressed thus: "Necessarily, any species with a type specimen contains its type specimen." The *de re* reading would be expressed: "Any species with a type specimen necessarily contains its type specimen". (LaPorte 2003, 586)

Laporte thinks that although the *de dicto* reading is true (2003, 587), the *de re* one is not, and this resolves the paradox. The first major concern of this paper is to argue that the *de dicto* reading, which I shall call "*Levine's Thesis*", is false. That is my conclusion *C1*.

LaPorte's response to Levine's alleged paradox was followed by several others: Matthew Haber (2012), Joeri Witteveen (2015), and Jerzy Brzozowski (2020). Haber argues that *Levine's Thesis* is false. Witteveen argues against Haber. Brzozowski defends Haber's position.

My argument for *C1* in section 3 appeals only to biology, with no mention of theories of reference. Indeed, I take the rejection of *Levine's Thesis* to be straightforwardly present in the words of biologists themselves. So why have some of these philosophers of biology accepted *Levine's Thesis* and all of them found the matter much more complicated? Answering that question is the other major concern of this paper. I shall argue that discussions of *Levine's Thesis*, whether for or against, have gone awry because of mistakes about language. One mistake is about the bearing of theories of reference on the assessment of a biological claim like *Levine's Thesis*. That is the subject of conclusion *C2*, argued in section 4. Another mistake is about reference itself. That is the subject of conclusion *C3*, argued in section 5. A final mistake is about the relation between linguistic decisions and the world. That is the subject of conclusion *C4*, argued in section 6. In sum, the engaging debate about *Levine's Thesis* has been misguided. In section 7, I consider some objections.

LaPorte's *de re* reading is not a major concern, but what about it? LaPorte thinks that it is false because of the possibility of the type specimen "never having been born" (2003, 587). I agree: no member is essential to a species. But he and Levine have another reason for thinking that the *de re* reading is false, one that LaPorte sets aside here (2003, 584). They both reject what LaPorte (1997) has aptly called "*Essential Membership*", the doctrine that an organism that belongs to a taxon

does so essentially. If no organism is essentially a member of its species, then no type specimen is. So, even if the actual type specimen for a species *is* born in another possible world, it might not be a member of that very species in that world. I must reject this reasoning because I have argued elsewhere (Devitt 2018b) *for Essential Membership*. Still, I agree that no type specimen of a species is necessarily a member of that species because, as we shall see in section 3, what counts against the *de dicto* reading (*Levine's Thesis*) counts also against the *de re* one.

2. The causal theory of reference and *Levine's Thesis*

Let us consider Levine's path to his Thesis. It starts with David Hull's "compelling account of the role of type specimens in the practice of taxonomy" (Levine 2001, 325), an account Hull offers in urging individualism and anti-essentialism about species.[2] Michael Ghiselin, who shares those views, is led to say: "As species are individuals, there is but one rigorous way to define their names: ostensively, in a manner analogous to a christening" (Ghiselin 1966, 209). Levine remarks: "It is interesting that Ghiselin's analogy to christening predates the literature on the Kripke–Putnam theory of reference (Levine 2001, 336, n. 3). And Levine notes that Hull was "quick to recognize" a connection between his view of type specimens and the Kripke–Putnam theory of reference:

> the importance [Hull] ascribes to the collection of type specimens in the ostensive naming of a species is strongly reminiscent of the role played by acts of baptism or dubbing in the Kripke–Putnam theory of rigid designation. (Ibid., 328)

Others noted this too (LaPorte 2003, 584; Haber 2012, 770; Witteveen 2015, 570; Brzozowski 2020, 2).[3]

---

[2] Ghiselin (1974) and David Hull (1978) take their view that species are *individuals* and not kinds to be an antidote to essentialism. I agree with those like Okasha (2002, 193–94) who think that this individualism is a red herring to the essentialism issue (Devitt 2008, 348).

[3] Devitt (2008, 2018a, 2018b) are among the papers cited by Brzozowski as offering "defenses of the causal-theoretical account of typification" (Brzozowski 2020, 7). This is very odd because there is no such defense in any of these papers. Indeed, their only mention of type specimens and the

Now I note first that the more usual, and much better, name for the Kripke–Putnam theory is "the causal theory of reference".[4] In any case, what was central and most novel about the Kripke–Putnam theory was not the appeal to dubbing, which we will consider in a moment, but the idea of epistemically undemanding *reference borrowing*: people who are very ignorant, even wrong, about the referent of a term, whether a proper name or a "natural kind" term, can nonetheless be competent users of the term simply in virtue of borrowing its reference from someone who was competent; there is a causal chain of such borrowings all the way back to the people who fixed the reference in a dubbing. This was a truly revolutionary idea. And Hull embraced that too:

> In rigid designation, a name is conferred in an initial baptismal act (possibly fictitious) and thereafter passed on in a link-to-link reference preserving chain. Regardless of the appropriateness of the Kripke–Putnam analysis in general, it accurately depicts the way in which systematists introduce the names of biological taxa. (Hull 1982, 491–492)

There was nothing novel, or particularly interesting, about drawing attention to dubbings as the typical way that proper names and some "natural kind" terms get their reference. Previous theorists of reference had not failed to notice the

---

causal theory of reference together is in a footnote sentence (Devitt 2018b, 39, n. 3) that concerns something else: the sentence foreshadows the conclusion that the causal theory does not imply *Levine's Thesis* (section 4).

[4] (I) Kripke (1980) carefully defined "rigid designator" for singular terms for the purpose of arguing that standard description theories of the reference-determining meaning of proper names are false. But, as quickly became apparent, this argument is easily avoided by a *description* (not causal) theory of rigid designation: a name's meaning is expressed by a *rigidified* description (Devitt and Sterelny 1999, 53–54). (II) The name "rigid designation" is particularly infelicitous for the Kripke–Putnam theory of "natural kind" terms. For, though Kripke extended his talk of "rigid designator" to general terms he did not provide a definition of its use for general terms. Just what the "rigidity" of such a term amounts to, or should amount to, is unclear, as quite a large literature shows; see, for example: LaPorte 2000; Schwartz 2002; Devitt 2005.

obvious fact that the names of many entities—babies, pets, ships, newly discovered animals and substances, and so on—typically acquire reference-determining meanings at baptisms and the like. But *what* meaning and reference was thus acquired in a dubbing, and *how*? *That was the issue.* The established "description theories" all assumed that the resulting reference was determined by descriptions that all competent with the new term associated with it. The major novelty of the Kripke–Putnam causal theory was, first, to reject that theory and, second, to emphasize that reference is fixed by dubbers *who then pass on the benefits of dubbings to others who may know little or nothing about the referent*. But what did the Kripke–Putnam theory tell us about that reference fixing in a dubbing? Not very much. Thus Kripke, discussing proper names in *Naming and Necessity*, talks briefly of "fixing a reference by description, or ostension" (Kripke 1980, 97). Howard Wettstein thinks fixing by description was Kripke's "paradigm" (Wettstein 2012, 115). Putnam talks of an "ostensive definition", but one accompanied by a description (Putnam 1975, 225–229): as he emphasized later, "descriptions play a key role: the original dubber or dubbers identify or have the capacity to identify what they are talking about by definite descriptions" (Putnam 2001, 496–97).

Indeed, it was hard then, and is hard now, for anyone to say much about what goes on in reference fixing. Ostension always struck me as the right way to go, but then what determines that a particular object is the object of ostension? There have been description theories of that too (Reichenbach 1947; Schiffer 1978). I favored a causal theory: reference is fixed in an object, directly or indirectly, by the causal link between a person and the object when it is the focus of that person's perception. This is what I call a "grounding" (Devitt 1974, 1981a).

So, on this view of reference fixing, the original users have their ability to designate Aristotle by "Aristotle" in virtue of a certain causal link to him and then we inherited this ability to designate him by reference borrowing. Even if one goes along with these old discussions of reference fixing, much is left unexplained, as I summarized in a recent update (Devitt 2015b). Still, those discussions did include a development

that is very relevant to *Levine's Thesis*, the idea of "multiple grounding". I will get to this in section 5.

Return to Hull and Levine. Given their individualism, they think that the name attached to a species by a type specimen is a *proper* name (Levine 2001, 329). They clearly reject the idea that the reference of that proper name is fixed by means of a description of the Aristotelian essence of the species. But then how do they think that reference *is* fixed? Levine has this to say:

> What allows such rigid designators to attach to their referents irrespective of the truth of any associated descriptions is that *they acquire their meanings in acts of dubbing or baptism*…. The similarity between the collection of type specimens, as understood by Hull, and such acts of baptism, should be evident. In the former, a biologist, in direct contact with a part of the target species (the specimen), attaches a name to a species without thereby proposing an Aristotelian definition. (Levine 2001, 328)

The theory of grounding that I have just described is clearly a "direct-contact" view of reference fixing and so it is not surprising that Levine (2001, 330–332) is sympathetic to it (and aware of some of its difficulties).

How do we get from this sort of causal theory to *Levine's Thesis*, "Necessarily, any species with a type specimen contains its type specimen"? The Thesis comes from the following view: "No matter in which species the type specimen is placed, its name goes with it" (Hull 1982, 484). Thus, the above-quoted passage, in which Hull likens the "rigid designation" theory's treatment of the "initial baptismal act" to the introduction of "the names of biological taxa", is followed by this:

> Both… require reference preservation. The respective terms cannot change their reference, although we can find out that we are mistaken about what we thought their reference was. (Hull 1982, 492)

This idea that the reference "cannot change" suggests to Levine that "the relation between a type specimen and the reference of its species name is… necessary" (Levine 2001, 334).

So Levine thinks that the causal theory applied to the species naming procedure implies *Levine's Thesis*. All his re-

spondents agree. Now, anyone who accepts this implication and favors the causal theory might well be led to embrace *Levine's Thesis*. Indeed, that is clearly the path of Levine and LaPorte; it seems also to be the path of Witteveen, as we shall see (sec 6.2). Yet is it really appropriate to embrace a biological thesis like Levine's on the basis of a theory of reference? I think not. Semantics should not be dictating to biology. Rather, semantics should answer to biology. This claim reflects the methodology of "putting metaphysics first" that I have argued for in a book of that name:

> We should approach epistemology and semantics from a metaphysical perspective rather than vice versa. We should do this because we know much more about the way the world is than we do about how we know about, or refer to, that world. (Devitt 2010, 2)

It follows that it is a mistake to use *any* semantic thesis to assess *any* biological thesis; the direction of assessment should be the reverse. Applying this to our particular issue yields another one of my conclusions, *C2*: *it is a mistake to use a theory of reference to assess Levine's Thesis*. My argument for this is in section 4.

Still we are interested in semantics as well as biology and so we do need a theory of reference that is compatible with the biological facts including, according to *C1*, the falsity of *Levine's Thesis*. In section 5, I shall argue that the causal theory is compatible *once we take account of multiple grounding*; for multiple grounding allows reference to change. So, I think that Levine and his respondents are wrong to accept the above implication: *the causal theory of reference does not imply Levine's Thesis*. This is my conclusion *C3*, to be argued in section 5.

I turn now to an evaluation of *Levine's Thesis*, an evaluation that will, of course, make no appeal to theories of reference.

3. The falsity of *Levine's Thesis*; the case for *C1*

Haber came up with an excellent example which has appropriately been at the center of the discussions of *Levine's Thesis* and will be at the center of mine:

> In the late 1990s a minor taxonomic scuffle arose over the endangered San Francisco Garter Snake (*Thamnophis sirtalis tetrataenia*, Cope in Yarrow 1875), and the common California Red-Sided Garter Snake (*Thamnophis sirtalis infernalis*, de Blainville 1835). Researchers discovered that *T. s. infernalis'* type specimen belonged to *T. s. tetrataenia* (Boundy and Rossman 1995; Barry et al. 1996). Typically in such cases the taxa would be re-named. The codes of taxonomic nomenclature are clear on this, with rules specifying just how to handle such cases, e.g., the principles of priority and typification (ICZN 1999, Art. 23, 61). In this case, though, a petition was submitted to the International Commission on Zoological Nomenclature (ICZN) requesting that the names be conserved for each taxon in question. The case was published (Barry and Jennings 1998), commentary solicited (Smith 1999), and a ruling issued (ICZN 1999): Opinion 1961 of the ICZN stated that a new type specimen had been designated for *T. s. infernalis*, thus conserving prevailing usage of the names. (Haber 2012, 767–8)

This example is about the type specimen of a *sub*species whereas *Levine's Thesis* is explicitly about species. Still what goes for the type specimen of a species goes for that of a subspecies. So we should take *Levine's Thesis* as being implicitly about subspecies too.

The 1835 type specimen, or holotype, for *T. s. infernalis* (originally *Coluber infernalis*) is held in a museum in Paris and catalogued as "MNHN 846" (Boundy and Rossman 1995). *Levine's Thesis* is:

> Necessarily, any species with a type specimen contains its type specimen.

Applying this to the subspecies *T. s. infernalis*, we get:

> Necessarily, *T. s. infernalis* contains its type specimen.

Does it? The resounding answer from experts is "No". The experts we need are those who know most about the type specimens of garter snakes, biologists, particularly taxonomists. We shall see that some think that the type specimen of *infernalis*, MNHN 846, *is not* a *T. s. infernalis* and others think that it *may well not be*. There is no sign of any expert thinking

that *it must be*. So, *Levine's Thesis* is false—conclusion *C1*—and there is no paradox.

It will help to show this if we identify two propositions that are entailed by the application of *Levine's Thesis* to this example. First, and most obviously:

> **HOLO**: MNHN 846, the type specimen for *T. s. infernalis*, is an *infernalis*.

Boundy and Rossman's claimed discovery that 846 is, in fact, from the snakes popularly known as San Francisco Peninsula garter snakes has not been contested. So let us assume it is so. Then, with that discovery, the application of *Levine's Thesis* entails that *T. s. infernalis* is (and always has been) the subspecies of those Peninsula snakes and not, as everyone has thought for decades, the subspecies of snakes popularly known as California coastal red-sided garter snakes. For, according to the discovery, 846, the type specimen of *T. s. infernalis*, is in the former subspecies not the latter. So:

> **INF** *T. s. infernalis* is the subspecies of San Francisco Peninsula garter snakes not the subspecies of California coastal red-sided garter snakes.

The very bad news for *Levine's Thesis* is simple: there is *no sign at all* of any expert endorsing either HOLO or INF and lots of signs of their not doing so.

Consider Boundy and Rossman 1995 on HOLO. They note that a 1941 review "restricted the name *infernalis* to the California coastal subspecies" and "revived the name *T. s. tetrataenia*" for "the San Francisco Peninsula populations" (Boundy and Rossman 1995, 236). As a result, at the time of their paper, as other biologists remark, "the taxonomy of the western subspecies of *Thamnophis sirtalis* has been resolved and well-accepted for 45 years" (Barry et al. 1996, 172). Boundy and Rossman have a detailed discussion of whether holotype MNHN 846 should be allocated to "either of the populations currently known as *T. s. infernalis* or *T. s. tetrataenia* or of an intermediate between the two" (Boundy and Rossman 1995, 237). They found that a certain

> combination of pattern elements on individual snakes is limited to the San Francisco Peninsula… within populations of typical

> *T. s. tetrataenia*. The geographic restriction of this pattern strongly indicates that the holotype of *C. infernalis* is assignable to those populations… The holotype belongs to a population(s) outside the geographic range and definition of *T. s. infernalis* as currently recognized. (Ibid., 238)

In other words, *MNHN 846 had been misidentified and is not an infernalis: HOLO is false*.

Now consider Barry and Jennings 1998. In their petition against Boundy and Rossman's proposal, they claim: "It is possible that the holotype of *T. s. infernalis* is a specimen of *T. s. tetrataenia*" (Barry and Jennings 1998, 224). In other words, *MNHN 846 might have been misidentified as an infernalis and HOLO might be false*. *Levine's Thesis* cannot allow this because it entails that 846 *cannot* be both a type specimen for *infernalis* and not an *infernalis*.

What about INF? Boundy and Rossman reject it also, but not so obviously. First, conspicuously, Boundy and Rossman do *not* say that, given their discovery about MNHN 846, we should embrace INF. Rather, their discussion of the "allocation" of 846 proceeds as if INF is not even under consideration. Thus, in making the comparisons that the allocation requires, they examined "approximately 200 specimens from within the range of *T. s. infernalis*". And their examination leads them to say that a certain marking on *Thamnophis sirtalis* "is reduced to irregular spotting, or re-placed by a broad, dark ventrolateral suffusion, in *T. s. infernalis*" (Boundy and Rossman 1995, 237). If INF were even a possibility given what Boundy and Rossman were revealing about 846, then rather than talk simply, as they do, of "*T. s. infernalis*", they should have said something like "the coastal snakes that *may have been wrongly identified as T. s. infernalis*". They are taking the falsity of INF for granted.

It's a similar story with Barry and Jennings (1998). As noted, they accept the possibility that MNHN 846 is not an *infernalis*. If *Levine's Thesis* were right, then this possibility would entail the possibility that INF is true. Barry and Jennings write as if this possibility has never occurred to them; Smith (1999), likewise. Thus, Barry and Jennings, after citing a large range of literature describing the Peninsula snakes as "*T. s. tetrataenia*", claim that "much of the same literature refers to *T. s. infernalis* as an allopatric form that does not occur

on the San Francisco Peninsula" (Barry and Jennings 1998, 225–226). There is no airing of the idea that this literature might be wrong because, given the facts about MNHN 846, *infernalis* might be *tetrataenia* and so INF might be true. Rather, Barry and Jennings presume INF is false.

Boundy and Rossman's discovery about MNHN 846 does not even raise the issue, for taxonomists, of whether the coastal snakes are *T. s. infernalis*. The issue actually raised by the discovery is quite different and is indicated by Haber: "typically in such cases the taxa would be re-named" (Haber 2012, 768). The issue raised is simply *which official names to use for the subspecies of Thamnophis sirtalis in the future*. Nothing more, nothing less. Should taxonomists follow the "default" (ibid., 777), according to the ICZN code, *re*naming *tetrataenia* "*infernalis*" and assigning a new name to *infernalis*, as Boundy and Rossman propose? Or should both subspecies retain their old names, as Barry and Jennings successfully petitioned? *All* parties see the issue raised by the discovery as simply over future names. Thus, for Boundy and Rossman, it is an issue of "nomenclatural changes" (1995, 238); for Barry and Jennings, one of "the rearrangement of the subspecies names" (1998, 226); for commentator Smith, one of "the stability of usage of these names" (1998, 72); finally, for the Commission, ICZN itself, in opinion 1961, the issue is

> the conservation of the subspecific name of *Thamnophis sirtalis infernalis* (Blainville, 1835) for the California red-sided garter snake from the Californian coast, and of *T. s. tetrataenia* (Cope in Yarrow, 1875) for the San Francisco garter snake from the San Francisco Peninsula… (ICZN 2000, 191)

This common understanding of the issue raised by MNHN 846 is at odds with INF and hence with *Levine's Thesis*. For, if INF were correct, there could be no question of *conserving* "*T. s. infernalis*" for the coastal snake since it would already be the name for the Peninsula snake not the coastal snake. And there could be no question of *re*naming the Peninsula subspecies "*T. s. infernalis*" because it would already have that name (even though nobody realized that it had!). It would have that name because MNHN 846 is the type specimen for *T. s. infernalis* and 846 is a Peninsula snake. The possibility that INF might be true is not even contemplated.

I conclude that Boundy and Rossman's uncontested discoveries about the type specimen, MNHN 846, are taken by those who know most about the type specimens of garter snakes not to imply either HOLO or INF. Taxonomy is rife with controversies but this is not one of them. So the experts reject *Levine's Thesis*. So we should too: conclusion C1.

I noted in section 1 that *Levine's Thesis* is LaPorte's *de dicto* reading of a claim that also has the following *de re* reading: "Any species with a type specimen necessarily contains its type specimen" (2003, p. 586). This reading is not a main concern but it is worth noting that the present discussion counts against that reading too. MNHN 846 was the type specimen for *T. s. infernalis*. Bounty and Rossman's uncontested discovery was that 846 had been misidentified and was not an *infernalis*. So the *de re* reading is false. (Since I endorse *Essential Membership* (Devitt 2018b), I think that 846 was necessarily a member of its species, *T. s. tetrataenia*. That is of course consistent with 846 being contingently a member of the species for which it was the type specimen, *T. s. infernalis*. So it does not create a new paradox.)

4. "But what about the theory of reference?"; the case for *C2*

In section 2 I foreshadowed the conclusion *C2*, that "it is a mistake to use a theory of reference to assess *Levine's Thesis*". Rather, the direction of assessment should be from biological facts to the theory of reference. So, my discussion of HOLO and INF has proceeded without appeal to a theory of reference. But *why* is it a mistake to make such an appeal? Why should we not follow Levine and others and argue as follows? "Our favorite theory of reference for biological kind terms, TR, tells us that, given the nature of MNHN 846, the name '*T. s. infernalis*' refers to the Peninsula snake not the coastal snake. So HOLO, INF, and *Levine's Thesis*, are true after all!" Problem: *Why believe TR*? Why not prefer a rival theory that tells us that "*T. s. infernalis*" refers to the coastal snake, or even to nothing at all? The traditional answer has been that TR *matches our referential intuitions*. Thus, TR predicts, time and again, that the reference of a biological kind term *E* in real or imagined situations is *X* and it just seems

intuitively to us philosophers that *E* does indeed refer to *X*. This methodology has been severely criticized in recent years. Many have argued that it is scientifically unsound and have insisted that theories of reference must be tested experimentally; see, for example, Machery et al. 2004; Machery et al 2009; Nichols et al 2016. Genoveva Martí (2009, 2012, 2014) and I (2011b, 2012a, 2012b, 2015a) have joined in the criticism and have gone on to argue that theories should be tested against *linguistic usage*.

This debate over methodology cannot of course be replayed here,[5] but I shall briefly apply the Martí–Devitt line to the present example. We should not accept any theory of reference for a term simply because its predictions conform to our intuitions about what the term refers to. Rather, we should test the theory against the usage of those competent with the term. So, TR needs to be tested against the usage of biologists particularly. Do these people *show by their usage* that they are referring to *X* by *E*? For example, does the taxonomists' use of "*T. s. infernalis*" show that they identify the Peninsula snake as its referent? Moral: *we need biologists opinion on the likes of INF in order to know whether TR is right*. Our only way now, perhaps ever, to determine whether a theory of reference for biological terms is right depends on our determination of the biological facts. The biologists' usage shows us that INF is false, as we have seen. So TR is false. That is the right direction of argument. No theory of reference has the evidential support to rule on INF and *Levine's Thesis*, contrary to what Levine and others presume. That is the case for *C2*.

Nonetheless, a theory of reference should be able to explain the linguistic usage demonstrated here, as anywhere. The causal theory mentioned in section 2, unlike TR, does explain that usage, once developed to include "multiple grounding".

[5] See Devitt and Porter 2021 for a summary of the literature and some examples of testing usage.

5. The causal theory of multiple grounding; the case for C3

As noted, my theory of "grounding" is a theory of the sort of reference fixing by "direct contact" that Hull and Levine favor. The most obvious examples of such groundings are the ceremonial dubbings that they mention. But there can be groundings without any such dubbings. Thus consider the naming of the cat Nana, discussed by Levine (2001, 330–1). This naming was by a dubbing but it could have been simply the result of usage: someone looking at Nana might have just said "Nana is a striking looking kitten" and thereby started the practice of calling the kitten "Nana". Nicknames are often introduced in this way. I recently summed up the theory of grounding as follows:

> What is it about all these situations that ground the name in a certain object? It is the causal-perceptual link between the first users of the name and the object named. What made it the case that this particular object got named in such a situation was its unique place in the causal nexus in the grounding situation. (Devitt 2015b, 114)

This leads straightforwardly to the theory of *multiple* grounding.

> It is important to note that this sort of situation will typically arise many times in the history of an object after it has been initially named: names are typically *multiply grounded* in their bearers. These other situations are ones where the name is used as a result of a direct perceptual confrontation with its bearer. The social ceremony of introduction provides the most obvious examples: someone says, "This is Nana", demonstrating the kitten in question. Remarks prompted by observation of an object provide many others: thus, observing Nana's behavior, someone says, "Nana is skittish tonight". Such remarks are likely to happen countless times during Nana's life. All these uses of a name ground it in its bearer just as effectively as does a dubbing because they involve just the same reference-fixing causal-perceptual links between name and bearer…. Dubbings and other first uses of a name do not bear all the burden of linking a name to the world. (Ibid., 114)

I used this idea of multiple grounding, together with Hartry Field's (1973) idea of partial reference, to explain cases of reference *confusion* (Devitt 1974, 200–203). Thus, consider Kripke's famous leaf-raking example: "Two people see Smith in the distance and mistake him for Jones" (Kripke 1979, 14). Suppose one person comments to the other, "Jones is raking the leaves". I argued that this use of "Jones" has a semantic-referent, Jones, but no determinate speaker-referent; both Jones and Smith are *partial* speaker-referents because the use is grounded in both (Devitt 1981b, 512–516; 2015b, 118–121). Later (Devitt 1981a, 138–152; 2015b, 121–124), I applied the ideas to cases of reference *change* including another famous example, Gareth Evans' "Madagascar" (Evans 1973). The story goes that Marco Polo, on the basis of a hearsay report of Malay sailors, mistakenly took the name of a portion of the African mainland, "Madagascar", as the name of the great African island. And that island is now, of course, the semantic-referent of "Madagascar". So "Madagascar" changed its reference. The explanation, in brief, is that the reference of a name changes from *x* to *y* when *the pattern of its groundings changes* from being in *x* to being in *y*.[6] This discussion is particularly relevant to *Levine' Thesis* if we go along with the individualist view that a species name is a proper name.[7]

Appeal to multiple grounding is also vital in explaining reference change in "natural kind" terms (Devitt 1981a, 190–5). Arthur Fine (1975, 22–6) criticized Putnam's causal theory

---

[6] Nonetheless, the mistaken idea that cases of reference change are "decisive against the Causal Theory of Names" (Evans 1973, 195) persists (Searle 1983; Sullivan 2010; Dickie 2011). Kripke's own response to "Madagascar" is in "Addenda" to *Naming and Necessity* (1980, 163). As I note (2015b, p. 123, n. 33), the grounding theory can be seen as an explanation of Kripke's admittedly brief proposal (but doubtless not one he would accept).

[7] So, it is odd that Levine does not mention this theory of reference change. He devotes much attention (2001, 330–332) to a discussion of "the qua problem" in chapter 4 of Devitt and Sterelny 1999, a textbook presentation of the causal theory of reference. That presentation includes the theory of reference change (75–76). Indeed, in the 1987 first edition which Levine uses, the theory of reference change immediately precedes the discussion of the qua problem.

of these terms on the ground that it makes it impossible for a term to change its reference: its reference is fixed by the original dubbing. Yet such scientific terms quite obviously often do change their reference. I pointed out (Devitt 1981a, 291–92, n. 1) that Putnam could easily add multiple grounding to his theory. And later he did: "As Devitt rightly observes, such terms are typically 'multiply grounded'" (Putnam 2001, 497). Reference change can then be explained, as it was with proper names, as a *change in the pattern of groundings* (Devitt 1981a, 192–5). This discussion would be particularly relevant to *Levine' Thesis* if we do not accept individualism as, it seems, most biologists do not.[8]

This explanation of reference change is not an *ad hoc* addition to the causal theory to solve problems. It is a straightforward corollary of the causal theory of groundings:

> Groundings fix designation. From the causal-perceptual account of groundings we get the likelihood of multiple groundings. From multiple groundings we get the possibility of confusion through misidentification. From confusion we get the possibility of designation change through change in the pattern of groundings. (Devitt 2015b, 123–124)

It is a truism among theorists of language that an expression gets its meaning and reference from conventions of usage. These conventions sometimes start with stipulations—dubbings are examples—but they mostly come from regular usage. However a convention is established, *even if by stipulation*, it can change through regular usage. (Think of the sad fate of "beg the question".) The above theory of groundings is an explanation of change for some sorts of words.

We now apply this theory to the names used to refer to Haber's garter snakes. An expression's conventional reference is typically established by regular usage. There was clear

---

[8] Ingo Brigandt claims that "most biologists and philosophers favor the idea that species are individuals rather than natural kinds" (2009, 77–8). Brigandt may be right about philosophers of biology—certainly the present debate provides evidence that he is—but a recent survey (Pušić et al 2017) shows he is quite wrong about biologists. The survey of 193 biologists from over 150 biology departments at universities in the US and the EU found that the position of individualism among biologists is "utterly marginal", only 2.94%.

consensus among taxonomists in the above debate that since 1951 there had been a stable usage of the name "*Thamnophis sirtalis infernalis*" to refer to California coastal red-sided garter snakes; see Barry and Jennings (1998), particularly. According to the causal theory this stability reflects *a pattern of groundings* of the name in those coastal snakes, a pattern of taxonomists (and others) using the name as a direct result of perceptual contact with those snakes. Doubtless in those decades, there were some groundings of the name in snakes of other kinds, particularly in MNHN 846, which is, after all, the type specimen for *T. s. infernalis* and yet is (we are assuming) a *tetrataenia*, not an *infernalis*. But these misidentifications pale into insignificance against the pattern of groundings in the coastal snake, *infernalis*. That pattern established and maintained the conventional use of the name "*Thamnophis sirtalis infernalis*" to refer to the coastal snake. And this is true whether we take the name to refer to an individual or to snakes of a certain kind.

According to Article 61 of the code, MNHN 846 should have provided "the objective standard of reference" (ICZN 1999) for "*Thamnophis sirtalis infernalis*": type specimens are supposed to stipulate a conventional usage. That is the thought behind Witteveen's claim: "If we baptize a specimen that belongs to some taxon as name-bearer, we thereby fix the name's reference to the taxon the specimen belongs to" (Witteveen 2015, p. 581). But the reference is thereby fixed only if all goes well for the stipulation. For, as just noted, stipulations can fail because expressions are not used as stipulated and different convention are established.[9] The consensus opinion about the usage of "*Thamnophis sirtalis infernalis*" shows that MNHN 846 is an example of such failure.

I emphasize that the Hullian idea that reference "cannot change" was *never* part of the Kripke–Putnam causal theory. Certainly the issue of reference change was not addressed in

[9] A corollary is that the following claims are false: "taxonomists had always known (with a priori certainty) that the *infernalis* type specimen belonged to the *infernalis* taxon" (Witteveen 2015, 582); "Type specimens… can be known a priori to belong to [their respective species]" (LaPorte 2003, p. 583). Knowledge of referential facts, indeed knowledge of semantic facts in general, is always empirical (Devitt 2011a; Salmon 2020).

the earliest presentations of the theory. Still it was in later ones. That is the case for C3: the causal theory of reference does not imply *Levine's Thesis*, as Levine and others think.

C2 identified the mistake by Levine and others of using a theory of reference to determine a biological thesis (sec. 4). That mistake is compounded by using a theory that does not accommodate reference change.

C3's rejection of the inference from the causal theory to *Levine's Thesis* has consequences for what Haber and Brzozowski say about reference. Given their acceptance of the inference, they take their arguments against *Levine's Thesis* to count against the causal theory (semantics appropriately answering to biology; sec. 2).[10] Thus, Haber thinks that his argument "suggests that rigid designation and causal theory of reference may be more fragile than supposed" (2012, 768).[11] The argument presents "a serious challenge to philosophical accounts of proper names, or perhaps their applicability to biological taxonomy" (ibid., 781). Brzozowski is led to the view that taxon names have their reference fixed by descriptions and are "descriptive names". He thinks that this "account of taxon names is able to better account for the uses and misuses of taxon names when compared to the causal view" (Brzozowski 2020, 23). *C3* undermines these criticisms of the causal theory.

6. Philosophical evaluations of *Levine's Thesis*

I turn now to the evaluation of *Levine's Thesis* by other philosophers. These evaluations include some claims which, from the perspective I have presented, are dead right. But they include others that are dead wrong. Thus, on the right side,

---

[10] If the rejection of *Levine's Thesis* poses a problem for the causal theory then, as LaPorte points out, it is "a general one": "it arises whether species are individuals or kinds, given the standard causal theory of reference" (LaPorte 2003, 586).

[11] Haber adds the following startlingly false claim: "Taxonomic theory is, in part, a theory of reference applied to biological nomenclature" (Haber 2012, 768). Taxonomic theory does specify *a practice for the stipulation of* a taxon name that will *cause* it to have a certain reference when all goes well, which it sometimes doesn't; but taxonomic theory is far from a *theory of* this reference.

Haber claims, contrary to HOLO, that "researchers discovered that *T. s. infernalis*' type specimen belonged to *T. s. tetrataenia*" (Haber 2012, 768) and goes on to reject *Levine's Thesis* and hence resolve the paradox. Brzozowski makes a similar claim (Brzozowski 2020, 10) and endorses Haber's rejection. Even Witteveen, who wrongly endorses *Levine's Thesis*, nonetheless apparently rejects INF in saying that Boundy and Rossman "discovered that taxonomists had been wrong about which taxon was [the *infernalis* type specimen's] taxon" (Witteveen 2015, 582).

But then there is the wrong side.

6.1 Haber; the case for C4

Haber's rejection of *Levine's Thesis* is strangely qualified: he thinks that the Thesis "only holds under idealized conditions" (Haber 2012, 782). This reflects a more serious problem: his reason for rejecting the Thesis confuses changing language with changing the world. This is the last of the "mistakes about language" that are a major concern of this paper.

My own reasons for rejecting *Levine's Thesis* arose from two related responses of taxonomist to the discovery about MNHN 846, the type specimen for the subspecies *T. s. infernalis*. These responses were contrary to what the Thesis demands. First, contrary to HOLO, these experts concluded that 846 had been, or might have been, misidentified as an *infernalis*, the California coastal red-sided garter snake; second, contrary to INF, these experts showed no sign of even entertaining the possibility that *infernalis* was not that coastal snake.

Now as noted in section 3, the discovery about MNHN 846 did demand a further response: taxonomists, particularly ICZN, had to make a decision about the future official names for the subspecies of *Thamnophis sirtalis*. *But the falsity of Levine's Thesis does not depend in any way on that decision about future usage*. Yet, as we shall see, Haber seems to think that it does. He seems to think that the Thesis *would be* true if ICZN always followed the code's "default" in such cases of misidentification, a default that would have been illustrated had ICZN accepted Boundy and Rossman's proposal that

tetrataenia be renamed "*infernalis*" and a new name be assigned to *infernalis*.

Abraham Lincoln is said to have once pointed out that a person's calling a donkey's tail a "leg" does not make it a leg. Similarly, the ICZN's calling the Peninsula snake "*T. s. infernalis*" would not have made it *T. s. infernalis*. It was a worldly fact that the Peninsula snake was not *T. s. infernalis*, no matter what decisions ICZN, or anyone, makes about how to use language in the future. Contrary to what postmodernists, and sadly many others, seem to think, languages do not make worlds. This is not the place to argue this large issue (but see, for example, Devitt 1997, 235–258; 2010, 99–136).

The key discussion in Haber begins nicely:

> That a specimen was preserved and identified prior to careful study of a particular taxon does not mitigate that the type specimen may be wrongly hypothesized to belong to that taxon. (Haber 2012, 779)

But then Haber goes on:

> In a default case, the species identity of the type specimen does not change, it still belongs to the species it designates. (*ibid*)

Had ICZN responded to the discoveries about MNHN 846 by deciding to follow the default it would have renamed *tetrataenia* "*infernalis*". This would have changed the status of 846: before such a decision, 846 does not belong to the subspecies for which it was a type specimen because it does not belong to *infernalis*; after the decision, it would have belonged to the subspecies for which it was a type specimen because it belongs to *tetrataenia*. But it would not have been in virtue of this decision that 846 kept its "species identity"! 846 was a *tetrataenia* (we are assuming) misidentified as an *infernalis*, showing *Levine's Thesis* to be false, *whatever linguistic decision anyone made about future usage*. Haber continues:

> On successful active petition… the type specimen… is reassigned to a new species, and no longer belongs to the species it formerly designated (though other specimens might). (*Ibid.*)

As Witteveen points out, Haber is arguing that the decision by ICZN to accept the petition of Barry and Jennings "entails that a type specimen got misidentified" (Witteveen 2015, 575).

Yet, what ICZN actually did was decide to conserve the subspecific *names* of both *T. S. infernalis* and *T. s. tetrataenia* (ICZN 2000, 191), rather than follow the default. This decision did not reassign MNHN 846 "to a new species" or entail that 846 had been misidentified. On the contrary, the decision is totally irrelevant to what (sub)species 846 belongs to. 846 had been misidentified as an *infernalis*, independent of any linguistic decision: to repeat, languages don't make worlds. Finally, contrary to what Haber claims (2012, 780), it is not because of that decision, rather than the default one, that the "*de dicto* necessity [*Levine's Thesis*] fails to hold". It fails simply because type specimens can be misidentified, as 846 illustrates. The "species identity" of any type specimen, like that of any organism, is constituted by its nature not by a linguistic decision of ICZN.

In sum, *it is a mistake to make any inferences about species identity, and hence about Levine's Thesis, from decisions about nomenclature*. This is my conclusion *C4*.

6.2 Witteveen

Witteveen claims to resolve Levine's paradox by arguing that "there is no sense in which type specimens belong contingently to the species they name" (Witteveen 2015, 571). Well, if my argument against *Levine's Thesis* is right then there is at least one such sense. Set that aside for a minute. According to LaPorte, there is another sense: the contingency that arises from the rejection of the *de re* necessity, "Any species with a type specimen necessarily contains its type specimen"? I argued that the misidentification of MNHN 846 provides one reason against this necessity (sec. 4). And LaPorte rightly points out that we should reject the necessity because of the possibility of the type specimen "never having been born" (LaPorte 2003, 587). Furthermore, he thinks, though I do not (sec. 1), that we should also reject this necessity because *Essential Membership* is false. So, there are *several* potential reasons for the contingency that comes from rejecting LaPorte's *de re* necessity. How does Witteveen resist all of them in claiming that that "there is no sense in which type specimens belong contingently to the species they name"? Briefly, by

confusing LaPorte's *de re* reading with his *de dicto* one (in a section called "Contingency confusion"!):

> Thus, it appears that in all possible worlds in which we find a species with a type specimen, it contains its type specimen. This means that the sentence "Any species with a type specimen necessarily contains its type specimen" is true after all. (Witteveen 2015, 576–7)

This is wrong. What appears to Witteveen to be so in his first sentence amounts to, "Necessarily any species with a type specimen contains its type specimen". This is LaPorte's *de dicto* reading, *Levine's Thesis*. This differs strikingly in the scope of its "necessarily" from what Witteveen takes the sentence to mean in his second sentence, namely, LaPorte's *de re* reading. And, the contingency we are considering is a rejection of the *de re* reading *not* the *de dicto* one. Witteveen has not addressed *that* "sense in which type specimens belong contingently to the species they name".

Return to Laporte's *de dicto* reading, *Levine's Thesis*. Witteveen's endorsement of this is, for our purposes, the key sense of contingency that he rejects. So, what is Witteveen's case for *Levine's Thesis*? It starts with criticism of Haber's case against. We have just rejected Haber's argument that the ICZN decision to accept Barry and Jennings' petition establishes that MNHN 846 was misidentified. Witteveen's criticisms are different. First, he claims:

> What Haber should have said" is that that ICZN decision "causes a specimen that formerly served as type specimen to stop belonging to the taxon for which it formerly anchored the taxon name. (Witteveen 2015, 580)

Now that decision *did* cause MNHN 846 to cease to be the type specimen of *infernalis*. But the decision *did not* cause 846 "to stop belonging to" *infernalis*: 846 never did belong. And no decision by ICZN could bear on the worldly fact of 846's subspecies membership; see conclusion *C4*. Witteveen's second criticism is better: he claims that the ICZN decision "does not show that *de dicto* necessity [*Levine's Thesis*] fails" (ibid., 581). No linguistic decision *could* show this. So Witteveen is right that Haber's case *against Levine's Thesis* fails. But what does Witteveen have to say *for Levine's Thesis*?

Only the passage we quoted and rejected earlier (sec. 5): "If we baptize a specimen that belongs to some taxon as name-bearer, we thereby fix the name's reference to the taxon the specimen belongs to" (ibid., 581). The problem was that attempts to stipulate usage can fail; reference can change (sec. 5). In any case, no thesis about language has the authority to settle a biological matter; see conclusion *C2*. To support *Levine's Thesis*, Witteveen needs to show that MNHN 846 was *not* misidentified as an *infernalis*, as taxonomists clearly think it (very likely) was. Witteveen has not done so.

6.3 Brzozowski

Brzozowski offers "a defense of Haber's (2012) position in response to Witteveen (2015)" (Brzozowski 2020, 4). Part of this defense is the rejection (ibid., 12) of a criticism of Haber that I have just emphatically endorsed: the charge that Haber takes *the ICZN decision* to entail that a type specimen got misidentified. In rejecting this criticism, Brzozowski points to a passage (Haber 2012, 778) like the one above that I labelled "on the right side". But the criticism is well-based in the cited passages "on the wrong side".

Brzozowski's discussion of this criticism, and his own remarks "on the right side" (Brzozowski 2020, 10), might suggest that he rightly thinks that the biological discovery that MNHN 846 had been misidentified alone shows that *Levine's Thesis* is false. But, in fact, he thinks that this discovery falsifies only a "metalinguistic" version of the thesis about "the reference of a species name" (ibid., 22). And this falsification depends on complicated semantic machinery, including the claim that names are descriptive (ibid., 14–23). This is a mistake: biology alone shows *Levine's Thesis* false. No semantics is needed; see conclusion *C2*.

I turn finally to some likely objections to my argument against *Levine's Thesis*.

7. Objections

I have a good basis for anticipating objections. For, the argument in this article has been presented before in a paper, "Type Specimens and Reference", that was rejected by two

journals on the basis of some thoughtful reports from reviewers.[12] I found the objections from two of these reviewers particularly interesting. The reviewers rightly think that issues about language have been center stage in the discussion of *Levine's Thesis* and *they insist that these issues continue to be*. Indeed, they find it incomprehensible that linguistic issues should not be put center stage. So, the reviewers are insisting on precisely what my paper argues is a very mistaken methodology. I shall develop my argument in this section in responding to the objections. It seems that this linguistic methodology is much more entrenched in this area of the philosophy of biology than I had supposed.

7.1 Reviewer R1 and codes of nomenclature

The objections from *R1* do not seem to be about language to begin with. *R1* claims that my

> bold argument would have been very interesting if it had been supported by convincing empirical evidence that taxonomists agree unanimously that it is not necessary for type specimens to belong to their species… I expected that the author would present evidence from questionnaires with vignettes of the kind that are frequently encountered in contemporary experimental philosophy (particularly in the area of semantics).

Section 3 presents fairly overwhelming evidence that *all* the taxonomists involved in the case of MNHN 846, and the international body ICZN itself, agree that 846, which is indubitably the type specimen for *Thamnophis sirtalis infernalis*, is, or at least might be, nonetheless a *T. s. tetrataenia*. What they agree on is inconsistent with *Levine's Thesis*. Now it is always good to have more evidence. So, we could see what taxonomists say about other cases of apparently misidentified type specimens. And we could indeed do some "experimental philosophy" on taxonomists. But if we do, we should not ask the taxonomists their opinion about whether it is "necessary for type specimens to belong to their species" (*Levine's Thesis*): that sort of question asked of taxonomists is

[12] The journals were *Biology and Philosophy* and *History and Philosophy of the Life Sciences*.

far too abstract and "philosophical" to provide good evidence. Rather, we should ask taxonomists about actual or imagined cases of apparently misidentified type specimens. This would provide good and direct evidence for or against *Levine's Thesis* of just the same sort as I provided. Indeed, we could present taxonomists with a vignette about MNHN 846 itself and ask them whether it is a *T. s. infernalis* or a *T. s. tetrataenia*; we could ask them about HOLO. But do we really need any of this extra evidence? Thus, given the *actual* discussion of 846 that I cited, we can surely be confident about their answer: 846, the type specimen for *T. s. infernalis* is, or at least might be, a *T. s. tetrataenia*.

This can't be *R1's* real worry about evidence and it soon becomes apparent that it isn't. The real worry is that the evidence that I provide from that actual discussion is "not viewed in the context of the debate" of Haber, Witteveen, and Brzozowski. What context is that? *A context that is largely about language*. Thus *R1* demands

> a close analysis of how this [rejection of *Levine's Thesis*] is supported by the wording of codes of nomenclature (ICZN, ICN and others) that taxonomists have devised and follow in their nomenclatural practices.

R1 charges that I do not "attend to the role of codes of nomenclature in taxonomic practice". *R1* finds this

> really quite baffling, since these codes—and their role in taxonomic practice—have been at the center of discussion in recent contributions to the "type specimen debate". By failing to consider the content and application of the codes in taxonomic practice, the author misses entirely what this type specimen debate has been about.

R1 is, of course, right that the debate over *Levine's Thesis* has centered on such linguistic matters. Indeed, I emphasized this at the very beginning of my discussion. So, I haven't *missed* it. Rather, I have emphatically *rejected* it: a "major concern" of the paper, and this article, is to argue that the debate has "gone awry because of mistakes about language" (sec.1).

How *might* a nomenclatural practice bear on *Levine's Thesis*? Here's a way. In section 4, I noted that a theory of reference, *TR*, could be brought to bear by telling us that,

"given the nature of MNHN 846, the name '*T. s. infernalis*' refers to the Peninsula snakes not the coastal snakes", thus supporting *Levine's Thesis*. Now suppose that *TR* tells us this about the name "*T. s. infernalis*" *because TR* takes the nomenclatural practice of stipulating a meaning for a taxon name via a type specimen to be what constitutes that reference to the Peninsula snakes. Then, clearly, the nomenclatural practice would provide evidence for *Levine's Thesis*. But, also clearly, the practice does so only if *TR* is *right* to give this role to the practice. And the problem is that *TR* is *not right* to. How do we know? Well, for "*T. s. infernalis*" to refer to the Peninsula snakes, there would have to be a convention of using it to so refer. That's a truism. And the usage by biologists shows that there is no such convention. Indeed, biologists had for decades been identifying the *coastal* snakes, not the Peninsula ones, as *T. s. infernalis*. *It is these identifications by biologists that provide the evidence for or against any theory of reference of "T. s. infernalis"* (Devitt and Porter 2021, 2023). Those identifications are what *TR* has to be tested against, and it fails.

But the moral of this tale is deeper. To assess *Levine's Thesis*, we need to know whether MNHN 846, the type specimen for *T. s. infernalis*, is a *T. s. infernalis* (HOLO). The deep moral is that it was a mistake to bring a theory of reference to bear on this question from the start (sec. 4). For, any theory of the reference of "*T. s. infernalis*" has to be tested against the term's usage. And the usage in question is that of taxonomists in identifying snakes as *T. s. infernalis* or not. So, to assess *Levine's Thesis*, we should simply check what biologists *do* identify as *T. s. infernalis* or not and skip the detour into the theory of reference. And that is what I did in section 3.

No application of a nomenclatural code constitutes the reference of "T. s. infernalis". That's a fact from the theory of language. There is no call for *R1* to be baffled by my inattention "to the role of codes of nomenclature in taxonomic practice". I attend to the only role played by these codes that is *relevant* to the reference of "*T. s. infernalis*". That role, I argue (sec. 5), is a *causal* not constitutive one. The application of a code is an obvious attempt to *stipulate* a term's reference, for important scientific purposes. And, of course, those attempts are mostly

successful: they establish a convention, thus causing the term to *have* that very reference. But, as the case of "*T. s. infernalis*" shows, sometimes stipulations fail because usage establishes different conventions. In sum, when all goes well for an authoritative body like ICZN, its stipulation that *E* is to refer to *S* will cause *E* to refer to *S*, but it never constitutes it so referring. That *E* refers to *S* is constituted by dispositions among *E*'s users (Devitt 2021, 75–81).

Despite the irrelevance of theories of reference to the assessment of *Levine's Thesis*, we do of course need a theory of reference that is compatible with the biological facts of the matter. I offered a causal theory of multiple grounding (sec.5). *R1* is not impressed, accusing me of failing "to see that taxonomists have agreed on the convention that only type designations 'ground' formal taxonomic names". Not guilty! Rather, what *R1* has failed to see is that *conventions agreed on may not be followed*; Geneva Conventions provide one example; "*T. s. infernalis*", another. *R1* continues: "One could in fact argue that one of the main purposes of the type method is to formally forbid 'multiple groundings' of taxon names". One could, but multiple groundings are a fact of linguistic life. So, it would be more plausible to argue that "one of the main purposes of the type method is to formally forbid" *groundings in any organism that is not in the same taxon as the type specimen*. That's plausible because the type method is a stipulation and stipulations indicate what people want. But, sadly, wanting something to be so, doesn't make it so. Thus, despite the Geneva Conventions, people got tortured. Similarly, despite the ICZN code, "*T. s. infernalis*" got multiply grounded in the coastal snake. So, the term *actually* refers to that snake. And *actual* reference matters to the theory of reference, not what the ICZN, or anyone, wants.

One might put my main point in response to *R1* as follows. The empirical methodology for the theory of reference, discussed in detail in the many works cited in section 4, and briefly described in that section and above, shows that the linguistic "context of the debate" over *Levine's Thesis* is mistaken. *R1* insists on that context without any recognition of that empirical methodology.

7.2 Reviewer R2 and the linguistic turn

R2 characterizes my methodology as follows: "we should simply ask experts (i.e., taxonomists) about whether *Levine's Thesis* holds". That's not quite right. My refutation of *Levine's Thesis* rests entirely on what taxonomists had to say about certain snakes, organisms that taxonomists know a lot about. The refutation does not rest at all on what taxonomists think about *Levine's Thesis*, a philosophical thesis that they might well find quite puzzling. In any case, *R2* objects:

> This methodology needs further motivation, since it is far from clear… that the taxonomists actually draw the conclusion that the Author claims they do. In particular, the Author will need to consider that the taxonomists he cites recognize the difference between the *usage* of names and their *valid* designation…. it is not evident that the taxonomists think that the valid name for a taxon can refer to a taxon that doesn't include the type for that name…. the Author appears to be holding the taxonomists to unreasonably high philosophical standards of precision in talking about naming and reference.… We can't expect taxonomists to neatly distinguish between these kinds of reference in their writings.

The opinions of taxonomists about snakes that I cite, including about type specimen MNHN 846, are inconsistent with *Levine's Thesis*. That is why we should reject *Levine's Thesis*. *R2* objects that we shouldn't reject it until we know what taxonomists think about *the names* of those snakes, until we have established that taxonomists have certain quite subtle *semantic* views. But, I responded to *R2*'s review, it was a central theme of my paper that views about language should *not* be used to assess a biological thesis like *Levine's Thesis*; see *C2* (sec .4) Any views about language, even ones held by expert semanticists, let alone by taxonomists, should not count against the views of expert taxonomists *about organisms*.

R2 was hugely unimpressed with this response, insisting that semantics *must* play a role. In particular *R2* finds it "really quite puzzling" how I "*could* think" that *Levine's Thesis* "is a purely biological thesis". For,

> a type specimen (a holotype or neotype) is nothing other than a specimen that serves as the bearer of a species name. So, we could rewrite [*Levine's Thesis*] as: "Necessarily, any species with a specimen that serves as the bearer of that species' name contains that specimen".[13] Is this a "purely biological" thesis? Surely not! It has semantics written all over it! Just consider a simple question this thesis invites: which is the species that the name-bearing specimen belongs to? Is it the name's semantic referent?

A consequence of *C2* is that this move to a semantic question is uncalled for and mistaken. Take our case of MNHN 846. Everyone agrees that 846 is the type specimen that serves as the bearer of the name for the species *T. s. infernalis*. Then *R2's* "simple question", applied to this case, is: "Does MNHN 846 belong to the semantic referent of '*T. s. infernalis*'?" But the question that should concern *Levine's Thesis* is not this partly semantic one but rather the entirely nonsemantic, "Is MNHN 846 a *T. s. infernalis*?" (cf. HOLO). And the resounding answer from people who know a lot about snakes, but probably very little about semantics, is "No (or probably not)". *That* is the judgment that refutes *Levine's Thesis*. *R2*'s insistence on bringing in semantics (without even addressing my argument that we should not) is very revealing of just how entrenched this "linguistic turn" is in this area of the philosophy of biology.

There is no sign that biologists involved in this case ever entertain *Levine's Thesis*, but they show by their practices that they reject it. So, they are not bothered by the problem allegedly posed by the Thesis. And they are right not to be. The alleged problem is a philosophical illusion, a misguided attempt by philosophers, driven by mistaken ideas about the relevance of views about language, to impose a problem on biology.

---

[13] *R2* actually proposed the following rewrite: "Necessarily, any species with a specimen that serves as the bearer of a species name belongs to the species of which it bears the name." But this must be a slip as it is clearly not a rewrite of *Levine's Thesis*. I have made corresponding adjustments in what follows the slip.

8. Conclusion

Levine (2001) sees a conflict between the contingency of species membership and a view of the role of type specimens that he takes from Hull: "*qua organism*, the type specimen belongs to its respective species contingently, while *qua type specimen*, it belongs necessarily"; he finds this "paradoxical" (ibid., 334). My concern has been with the thesis about type specimens which, following LaPorte, I take to be the *de dicto* necessity, "Necessarily, any species with a type specimen contains its type specimen" (LaPorte 2003, 586). I called this "*Levine's Thesis*". I have used Haber's lovely example of MNHN 846, the type specimen for *Thamnophis sirtalis infernalis*, to argue for conclusion *C1*: *Levine's Thesis* is false (sec. 3). For, the uncontested discovery by two taxonomists, Boundy and Rossman (1995), is that 846 is not a *T. s. infernalis* but a *T. s. tetrataenia*.

The alleged paradox has led to papers not only one from LaPorte but also from Haber (2012), Witteveen (2015), and Brzozowski (2020). My argument for *C1* appealed only to biology, with no mention of theories of language. In this respect it differs from other arguments about *Levine's Thesis*, whether for it or against it. A major concern of this paper has been to show that these arguments have gone awry because of mistakes about language.

First, Levine's path to *Levine's Thesis* rests on a causal theory of reference which he takes from Kripke and Putnam. My conclusion *C2* was that it was a mistake for Levine to use a theory of reference to assess *Levine's Thesis*; the direction of assessment should be from biological facts to the theory of reference (sec. 4). This criticism applied also to LaPorte's and Witteveen's arguments for *Levine's Thesis* and to Brzozowski's argument against.

Still we are interested in semantics as well as biology and so need a theory of reference compatible with the biological facts. So, we need a theory that does not imply *Levine's Thesis*. I argued against the received view that the causal theory does imply this: that's my conclusion C3 (sec. 5). A causal theory that includes multiple groundings can explain reference change and accommodate the falsity of *Levine's Thesis*.

The final mistake is about the relation between linguistic decisions and the world (sec.6). Haber rightly rejects *Levine's Thesis*, but he does so for the wrong reason. In response to Barry and Jennings' (1998) petition about the MNHN 846 discovery, ICZN (2000) decided to conserve the subspecific names of both *T. S. infernalis* and *T. s. tetrataenia*. Haber thinks that it was this decision that made it the case that 846 had been misidentified as an *infernalis*, hence establishing the falsity of *Levine's Thesis*. Witteveen, who accepts *Levine's Thesis*, has a different view of what that decision achieved: it caused 846 to stop belonging to *infernalis*. It followed from my conclusion *C4* that both these views are wrong: it is a mistake to make any inferences about species identity, and hence about *Levine's Thesis*, from decisions about nomenclature; changing languages does not change worlds. Whether or not 846 is an *infernalis* or a *tetrataenia* and hence has been misidentified is a biological fact that does not depend in any way on a linguistic decision.

I ended my discussion by responding to some objections taken from a couple of unfavorable reviews (sec.7). These reviewers wrongly insist on putting linguistic issues center stage in discussing *Levine's Thesis*, despite my argument that this is a mistake (*C2*).

Levine's Thesis is false. So, there would be no paradox even if *Essential Membership* were not true. But it is true (Devitt 2018b).[14] This does not yield a new paradox. According to *Essential Membership*, MNHN 846 is necessarily a member of its species, *T. s. tetrataenia*. That is quite consistent with the falsity of *Levine's Thesis*: it is consistent with 846 not necessarily being a member of *T. s. infernalis*, the species for which it is a type specimen; indeed, with it not being a member of that species at all.[15]

Graduate Center, City University of New York

[14] In the version of this paper that appears as ch. 5 in my book, *Biological Essentialism*, the "not" in this sentence was mistakenly moved to the next sentence leading to the false claim that *Essential Membership* "is not true" (Devitt 2023, 156).

References

Barry, S. J. and M. R. Jennings (1998). *Coluber infernalis* Blainville, 1835 and *Eutaenia sirtalis tetrataenia* Cope in Yarrow, 1875 (Currently *Thamnophis sirtalis infernalis* and *T s tetrataenia*; reptilia, squamata): Proposed Conservation of the Subspecific Names by the Designation of a Neotype for *T s infernalis*." *Bulletin of Zoological Nomenclature* 55(4), 224–228.

Barry, S. J., M. R. Jennings, and H. M. Smith (1996). "Current Subspecific Names for Western *Thamnophis sirtalis*." Herpetological Review 27(4), 172–173.

Boundy, J. and D. A. Rossman (1995). "Allocation and Status of the Garter Snake Names *Coluber infernalis* Blainville, *Eutaenia sirtalis tetrataenia* Cope and *Eutaenia imperialis* Coues and Yarrow." *Copeia* 1995(1), 236–240.

Brigandt, I. (2009). "Natural Kinds in Evolution and Systematics: Metaphysical and Epistemological Considerations." *Acta Biotheoretica* 57, 77–97. https//:doi.org/101007/s10441-008-9056-7

Brzozowski, J. A. (2020). "Biological Taxon Names Are Descriptive Names." *History and Philosophy of the Life Sciences* 42, 29. https://doi.org/101007/s40656-020-00322-1

Cope, E. D. in Yarrow HC (1875). "Report upon the Collections of Batrachians and Reptiles made in Portions of Nevada, Utah, California, Colorado, New Mexico, and Arizona, during the years 1871, 1872, 1873 and 1874", 509–584. In: Engineer Dept, USA (ed.), Report upon Geographical and Geological Explorations and Surveys West of the One Hundredth Meridian," vol. 5 (*Zoology*), part 4, 546.

de Blainville, H. (1835). "Description de quelques espèces de reptiles de californie: précédée de l'analyse d'un système général d'erpétologie et d'amphibiologie." *Nouvelles Annales du Muséum d'Histoire Naturelle* 3(4), 291.

Devitt, M. (1974). "Singular Terms." *Journal of Philosophy* 71, 183–205.

Devitt, M. (1981a). *Designation*. New York: Columbia University Press.

Devitt, M. (1981b). "Donnellan's Distinction." In P. A. French, T. E. Uehling Jr., and H. K. Wettstein (eds.), *Midwest Studies in Philosophy, Volume VI: The Foundations of Analytic Philosophy*. Minneapolis: University of Minnesota Press, 511–524.

Devitt, M. (1997). *Realism and Truth*, 2nd edn. Princeton: Princeton University Press.

Devitt, M. (2005). "Rigid Application." *Philosophical Studies* 125, 139–165.

Devitt, M. (2008). "Resurrecting Biological Essentialism." *Philosophy of Science* 75, 344–382. Reprinted in Devitt (2010) with some substantive additional footnotes.

Devitt, M. (2010). *Putting Metaphysics First: Essays on Metaphysics and Epistemology.* Oxford: Oxford University Press.

Devitt, M. (2011a). "No Place for the A Priori." In M. J. Shaffer and M. L. Veber (eds.), *What Place for the A Priori?* Chicago and La Salle: Open Court Publishing Company, 9–32. Reprinted in Devitt 2010.

Devitt, M. (2011b). "Experimental Semantics." *Philosophy and Phenomenological Research* 82, 418–435. https://doi.org/ppr201182222

Devitt, M. (2012a). "Whither Experimental Semantics?" *Theoria* 27, 5–36.

Devitt, M. (2012b). "Semantic Epistemology: Response to Machery." *Theoria* 27, 229–233. https://doi.org/ theoria20122711101387/theoria6225

Devitt, M. (2015a). "Testing Theories of Reference." In J. Haukioja (ed.), *Advances in Experimental Philosophy of Language*. London: Bloomsbury Academic, 31–63.

Devitt, M. (2015b). "Should Proper Names Still Seem So Problematic?" In A. Bianchi (ed.), *On Reference*. Oxford: Oxford University Press, 108–143.

Devitt, M. (2018a). "Historical Biological Essentialism." *Studies in History and Philosophy of Biological and Biomedical Sciences* 71, 1–7. https://doi.org/101016/jshpsc201805004

Devitt, M. (2018b). "Individual Essentialism in Biology." *Biology and Philosophy* 33, 1–22. https://doi.org/101007/s10539-018-9651-1

Devitt, M. (2023). *Biological Essentialism*. Oxford: Oxford University Press.

Devitt, M. and Porter, B. C. (2021). "Testing the Reference of Biological Kind Terms." *Cognitive Science* 45. doi.org/10.1111/cogs.12979

Devitt, M. and Porter, B. C. (2023). "Two Sorts of Biological Kind Terms: The Cases of 'Rice' and 'Rio de Janeiro Myrtle'". *Philosophy and Phenomenological Research*. doi: 10.1111/phpr.12979.

Devitt, M. and K. Sterelny (1999). *Language and Reality: An Introduction to the Philosophy of Language*, 2nd edn. (1st edn 1987.) Oxford: Blackwell Publishers.

Dickie, I. (2011). "How Proper Names Refer." *Proceedings of the Aristotelian Society* Vol. 101, 43–78.

Evans, G. (1973). "The Causal Theory of Names." *Proceedings of the Aristotelian Society*, Supplementary Volume 47, 187–208.

Field, H. (1973). "Theory Change and the Indeterminacy of Reference." *Journal of Philosophy* 70, 462–81.

Fine, A. (1975). "How to Compare Theories: Reference and Change." *Noûs* 9, 17–32.

Ghiselin, M. (1966). "On Psychologism in the Logic of Taxonomic Controversies." *Systematic Zoology* 26, 207–215.

Haber, M. H. (2012). "How to Misidentify a Type Specimen." *Biology and Philosophy* 27, 767–784.

Hull, D. (1982). "Exemplars and Scientific Change." In P. D. Asquith and T. Nickles (eds.), *PSA 1982, Vol II*. East Lansing: Philosophy of Science Association, 479–503.

ICZN (International Commission on Zoological Nomenclature) (1999). International Code of Zoological Nomenclature, 4th edn. The International Trust for Zoological Nomenclature 1999. http://www.nhmacuk/hosted-sites/iczn/code/. Accessed 20 April 2012.

ICZN (International Commission on Zoological Nomenclature) (2000). Opinion 1961: *Coluber infernalis* Blainville, 1835 and *Eutaenia sirtalis tetrataenia* Cope in Yarrow, 1875 (Currently *Thamnophis sirtalis infernalis* and *T s tetrataenia*; reptilia, serpentes): Subspecific Names Conserved by the Designation of a Neotype for *T s infernalis*." *Bulletin of Zoological Nomenclature* 57(3), 191–192.

Kripke, S. A. (1979). "Speaker's Reference and Semantic Reference." In P. A. French, T. E. Uehling Jr., and H. K. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, 6–27.

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

LaPorte, J. (1997). "Essential Membership." *Philosophy of Science* 64, 96–112.

LaPorte, J. (2000). "Rigidity and Kind." *Philosophical Studies* 97, 293–316.

LaPorte, J. (2003). "Does a Type Specimen Necessarily or Contingently Belong to Its Species?" *Biology and Philosophy* 18, 583–588.

Levine, A. (2001). "Individualism, Type Specimens, and the Scrutability of Species Membership." *Biology and Philosophy* 16, 325–338.

Machery, E., R. Mallon, S. Nichols, and S. P. Stich (2004). "Semantics, Cross-Cultural Style." *Cognition* 92, 1–12 https://doi.org/101016/jcognition200310003

Machery, E., C. Y. Olivola, and M. de Blanc (2009). "Linguistic and Metalinguistic Intuitions in the Philosophy of Language." *Analysis* 69, 689–694. https://doi.org/101093/analys/anp095

Martí, G. (2009). "Against Semantic Multi-culturalism." *Analysis* 69, 42–48. https://doi.org/101093/analys/ann007

Martí, G. (2012). "Empirical Data and the Theory of Reference." In W. P. Kabasenche, M. O'Rourke, and M. H. Slater (eds.), *Reference and Referring: Topics in Contemporary Philosophy*. Cambridge, MA: MIT Press, 62–76.

Martí, G. (2014). "Reference and Experimental Semantics." In E. Machery and E. O'Neill (eds.), *Current Controversies in Experimental Philosophy*. New York: Routledge, 17–26.

Nichols, S., N. A. Pinillos, and R. Mallon (2016). "Ambiguous Reference." *Mind* 125, 145–175. https://doi.org/101093/mind/fzv196

Pušić, B., D. Franjević, and P. Gregorić (2017). "What Do Biologists Make of the Species Problem?" *Acta Biotheoretica* 65(3), 179–209. https:// doi.org/01007/s10441-017-9311-x

Putnam, H. (1975). *Philosophical Papers, Vol. 2*: *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Putnam, H. (2001). "Reply to Devitt." *Revue Internationale de Philosophie* 208, 495–502.

Reichenbach, H. (1947). *Elements of Logic*. London: Macmillan.

Salmon, N. (2020). "Naming and Non-necessity." In A. Bianchi (ed.), *Language from a Naturalistic Perspective: Themes from Michael Devitt*. Cham: Springer, 237–248.

Schiffer, S. (1978). "The Basis of Reference." *Erkenntnis* 13, 171–206.

Schwartz, S. P. (2002). "Kinds, General Terms, and Rigidity." *Philosophical Studies* 109, 265–277.

Searle J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Smith, H. M. (1999). "Comment on the Proposed Conservation of *Coluber infernalis* Blainville, 1835 and *Eutaenia sirtalis tetrataenia* Cope in Yarrow, 1875 (Currently *Thamnophis sirtalis infernalis* and *T s tetrataenia*; Reptilia, Squamata): Proposed Conservation of the Subspecific Names by the Designation of a Neotype for *T s infernalis*." *Bulletin of Zoological Nomenclature* 56(1), 71–72.

Sullivan, A. (2010). "Millian Externalism." In R. Jeshion (ed.), *New Essays on Singular Thought*. Oxford: Oxford University Press, 246–269.

Wettstein, H. K. (2012). "On Referents and Reference Fixing." In R. Schantz (ed.), *Prospects for Meaning*. Berlin: De Gruyter, 107–18.

Witteveen, J. (2015). "Naming and Contingency: The Type Method of Biological Taxonomy." *Biology and Philosophy* 30, 569–586. https://doi.org/101007/s10539-014-9459-6

Conceptual Engineering for Externalists

JUSSI HAUKIOJA

1. Introduction

Conceptual engineering, the project of improving our concepts and other representational devices, has received considerable attention and enthusiasm in recent philosophy. Many of our concepts, it is argued, do not make the distinctions we ideally *should* make, in order to succeed in our (political, philosophical, ethical, practical, and so on) aims. We should therefore strive to revise these concepts. Examples of philosophically interesting concepts that have been argued to stand in need of improvement are those of *truth, belief, race, woman, knowledge,* etc. For simplicity, I will here follow many others in the debates and understand conceptual engineering as primarily consisting in intentionally changing the *intensions* of our *words*.[1] Intensions are here understood as functions from possible worlds to extensions, or less technically as criteria for belonging in extensions. Conceptual engineering thereby involves changing the extensions of our words, not by manufacturing or destroying things, but by changing what it takes to belong in the extensions.

Many theoretically interesting problems connected to conceptual engineering have been pointed out and discussed. My focus here will be on possible problems with combining conceptual engineering and *semantic externalism*, a widely held view regarding how linguistic meaning is determined. Semantic externalists hold that the meanings of our terms (or, at least some of them) are at least partly dependent on exter-

[1] I am understanding conceptual engineering to operate on semantic meaning. This assumption is widespread, but not universally accepted: see Pinder (2021) for a defense of conceptual engineering as primarily concerned with speaker meaning.

nal matters of fact. Semantic externalism comes in many flavors—my main focus here will be on the two most discussed and widely accepted externalist views, natural kind externalism and social externalism. Natural kind externalism is typically motivated by thought experiments such as Putnam's Twin Earth (Putnam 1975), and holds that the intensions of natural kind terms, such as "water," are partly determined by features of our natural environment, such as the chemical structure of the local watery stuff (that is, the tasteless, colorless substance predominately found in lakes, rivers, taps, and so on). Social externalism, on the other hand, holds that the intensions of many (possibly all) terms are partly determined by facts concerning other speakers (e.g., Burge 1979). One widely held social externalist view, which we will come back to below, claims that the meanings (and thereby intensions) of some terms are determined by *experts* who can make the appropriate distinctions, while the rest of us use the relevant terms with the same meaning as the experts, because we semantically *defer* to these experts (Putnam 1975).

2. The problem

It is not hard to see how a potential tension between conceptual engineering and semantic externalism arises. As noted above, conceptual engineering involves intentionally changing the intensions of our terms. According to semantic externalism, on the other hand, the intensions of our terms can depend on external matters of fact such as chemical structures, and/or the beliefs and linguistic behavior of experts. Typically, we have little or no control over such facts:

> […] effecting conceptual change looks comparatively easy from an internalist perspective. We can revise, eliminate, or replace our concepts without worrying about what the experts are up to, or what happens to be coming out of our taps. From the externalist's point of view, however, conceptual revolution takes a village, or a long trip to Twin Earth. (Burgess and Plunkett 2013, 1096)

Steffen Koch spells out the problem as follows:

> (1) SE [semantic externalism] is true about many terms in our language, and in particular those terms typically in the focus of practitioners of CE [conceptual engineering].
>
> (2) If SE is true about a given term *t*, then it is not within our control to change the meaning of *t*.
>
> (3) If it is not within our control to change the meaning of *t*, CE is not applicable to *t*.
>
> (4) Therefore, CE is not applicable to many terms of our language, and in particular it is not applicable to those terms typically in the focus of practitioners of CE. (Koch 2021, 330–331)

Note, however, that at least some *social* externalist views appear to be relatively unproblematic, with respect to conceptual engineering. In particular, the kind of view mentioned above, which holds that the intension of term *t* is determined by the relevant experts' usage (to which non-experts defer), does not pose any special problems for conceptual engineering. Depending on how the experts' usage determines the intension, we get two main kinds of case. In the first, the intension is determined by the properties/descriptions/definitions associated with *t* by the experts. When this is the case, conceptual engineering may of course be challenging for various pragmatic or social reasons, but there is no deep conceptual problem: if the experts agree to change the definition (etc.) that they associate with the term, while the rest of us go on deferring to them, the intension of the term has changed. Arguably, this is exactly what happened when the International Astronomical Union changed the definition of "planet" in 2005. If, on the other hand, the intension of a term is determined by the experts' causal interactions with the kind/phenomenon in question (as it arguably is in Putnam's influential examples of "elm" and "beech"), social externalism does not cause any *extra* problems: whatever difficulties there are, in engineering the meanings of such terms, stem from them being natural kind terms.[2] Accordingly, my main focus below will be on natural kind externalism.

---

[2] Another kind of social externalism might hold that the intensions of some or all terms are determined socially, but without deference to a par-

Koch's solution, as mine, is to reject (2). I will discuss Koch's view, as well as present my own, in the next section. But it should be noted that neither (1) nor (3) is obviously true: one could also react to the problem by denying one of them. As for (1), none of the examples mentioned in the introduction are obvious examples of natural kind concepts, although some would hold that, e.g., *knowledge* is a natural kind. However, there's no obvious reason why some of our natural kind concepts might not stand in need of improvement: denying (1) would seriously limit the scope of conceptual engineering.[3] Cappelen (2018) is plausibly read as denying (3). I will not discuss his positive view here (but I will, in Section 4, comment on his objection to the kind of view I propose below)—here it is enough to note that his view, too, is unduly pessimistic about the scope and prospects of conceptual engineering, if (2) can be rejected.

3. Rejecting (2): semantic externalism and meaning control

3.1 Koch's proposal

Koch's solution to the problem starts with the observation that all main variants of semantic externalism already allow for reference change (where this is a result of a change in an externally determined intension, rather than merely a non-semantic change in the world, causing changes in extensions while the relevant intension remains unchanged). This is ap-

---

ticular set of experts. It might, for example, be held that individual speakers defer to how the *majority* of other (competent) speakers in their linguistic community use said terms. Such a view would, for example, seem to fit well with Burge's (1979) discussion of terms like "sofa," although Burge does not explicitly commit himself to it. I assume that such social externalist views would not cause principled problems for conceptual engineering—in the case of such terms, conceptual engineering would merely require changing the speech patterns of the majority of speakers in a community—but I will not discuss this issue in detail here.

[3] But see Haslanger (2006) for the view that something like natural kind externalism applies much more widely than often assumed, in particular that it applies to the social kind terms often focused on in discussions of conceptual engineering.

parent, for example, in discussions of so-called *slow switching* cases, where a speaker is transported to a new environment containing a natural kind superficially similar to a kind the speaker was previously familiar with, but with a different underlying structure (Burge 1988). It is generally assumed by externalists, for example, that although an Earthling's early tokens of "water" after arrival on Twin Earth would only denote H_2O, were the speaker to remain on Twin Earth and keep calling XYZ "water," eventually her tokens of "water" would change their meaning, and their extension would then include XYZ. These two cases are discussed in some detail by Koch, in his thought experiments of *Young-Mary* and *Old-Mary*, respectively (Koch 2021, 336–337).

Different externalist theories would account for such changes in different ways (cf. Evans 1973; Devitt 1981). For example, according to Evans's theory, which Koch chooses as his illustrative example, a natural kind term such as "water" refers, roughly, to the substance that is the causal source of the body of information that the speaker associates with the term. For an Earthian speaker who has recently been transported to Twin Earth, H_2O is still the main causal source of the information she associates with "water," but after a sufficient time, XYZ will have taken its place, as now most of the information the speaker associates with "water" will have XYZ as its causal source. When that has happened, the meaning of "water," as used by the speaker, has changed. The details of the explanation are not crucial here—what matters for Koch's view is that we *already* think semantic externalism (and natural kind externalism in particular) is consistent with a term's intension changing over time. Provided we have some account of when and how intensions change, what would then stop us from effecting such changes intentionally?

Externalism is then, Koch argues, compatible with what he calls *collective long-range control:* by collectively adopting new ways of speaking about (*e.g.*) natural kinds, we can intentionally bring about meaning change over time (assuming standard externalist views of meaning change are at least roughly along the right lines):

> Many people start using the term in question *as if it had the new reference*; eventually, this will add pieces to the body of infor-

> mation we associate with the term that have the new object or kind as their causal source. […] Thus, little by little, the term will shift from the old reference to the new one [...]" (Koch 2021, 343).[4]

I fully agree with Koch that collective decisions regarding language use can result in intentional meaning change, even if externalism is true of the relevant expressions. However, I disagree with Koch's explanation of *how* such collective decisions can change meanings. In the next section, I will argue that meaning change is in fact, in a sense, easier to accomplish than Koch would allow for, even of externalism is true.

Moreover, I am not convinced that slow switching cases provide us with a good model for explaining intentional meaning change. In slow switching cases, there is by hypothesis no change in the communicative behavior of the Earth-to-Twin-Earth traveler: she continues to apply the term in the same way as before, based on how the situation appears to her. Yet, we are inclined to say that at some point the truth value of her utterances of, say, "there is water in that lake" (pointing to a lake on Twin Earth), will change. The meaning change is not caused by changes in how the speaker is disposed to apply the term, but rather by changes in the environment, of which the relevant speakers are moreover typically assumed to be unaware of. In the kinds of conceptual engineering projects that Koch envisages, by contrast, the environment remains unchanged in the relevant respects: the supposed change in meaning is a result of changes in how the speakers apply the term in question, based on how the situation appears to them. Given this asymmetry, it is not at all obvious that the rate at which the meaning change occurs is similar in in the two cases. In the next section, I will argue that there is good reason to think that the two kinds of situation are crucially different.

---

[4] Based on the quotation, it might seem as if Koch takes the reference shift to be gradual. However, I think it is charitable to interpret him as claiming the reference shift to be instantaneous: what is gradual is, rather, the process of the *preconditions* of reference shift gradually building up. I am grateful to an anonymous reviewer for pointing this out to me.

3.2 An easier way to reject (2)

Let us start with a thought experiment. Suppose that, sometime in the future, humans discover Twin Earth, which is just as Putnam (1975) famously imagined it to be. Suppose, moreover, that the distance between Earth and Twin Earth is manageable for the technology then available, and we begin frequent travel between Earth and Twin Earth. The chemical difference between the planets is by then well known, of course, and at first speakers take great care to keep track of which planet they are on, and call the liquid they are dealing with either "water," or "twin water," accordingly. However, as the interplanetary travel goes on, this gradually becomes perceived as an unnecessary cognitive burden on speakers—the difference has no impact on their daily lives, after all. And the Twin Earthlings will of course go on calling water "twin water" and twin water "water," just as meticulously, making things even more confusing. Sooner or later, the speakers (of both English and Twin English) decide that life would be a lot easier if everyone just used "water" to talk about watery stuff—*any* clear, odorless, thirst-quenching liquid that fills lakes and rivers, comes out of taps, and so on. This suggestion gains wide acceptance, the populations of the two planets are informed, and everyone conforms to the new usage. ("H_2O" and "XYZ," or some newly introduced terms, are then used in contexts where the chemical composition does make a difference.)

In this imagined scenario, all speakers (of both English and Twin English) switch from applying "water" on the basis of (assumed) sharing of chemical structure with the watery stuff on their respective home planets, to applying it merely on the basis of manifest properties.[5] This fits Koch's description of

[5] It might be objected that the change imagined here is so dramatic that it amounts to a change of *topic* rather than a meaning change that is consistent with speakers still discussing the same topic. The question of topic continuity is another contested issue connected to conceptual engineering (see, *e.g.,* Cappelen 2018; Sawyer 2018). A discussion of topic continuity falls outside the scope of the present paper: if it turns out that there is no topic continuity in the case imagined here, a structurally similar thought experiment could be formulated, where the change in meaning is less dramatic (and consistent with topic continuity).

how we effect "long-term collective control" over the meanings of our terms: speakers "start using the term in question *as if it had the new reference*." On his view, then, the extension change would take place only after a substantial delay (when the new usage has become the main causal source of information, or the new usage has been in place long enough for multiple grounding to have taken place—the details will depend on our preferred externalist theory of reference change). But can this be right? Remember that the change in the speakers' speech patterns is imagined to be more or less instantaneous: all Earthlings and Twin Earthlings decide to use "water" to refer to all watery stuff, interpret each others' use of the term in the same way, and communicate perfectly using the term. Yet, according to Koch, we should say that the Earthlings' "water" continues to refer only to H_2O, and the Twin Earthlings' "water" to XYZ, for a substantial amount of time, and that speakers utter systematic falsehoods in a substantial range of cases, until at some point in the future the semantic facts click in place. Moreover, when the semantic facts *do* click in place, the only thing that really changes is the truth values of the sentences uttered by the speakers: all the changes in the speakers' communicative behavior took place long before this.

This should strike us as an odd consequence. According to our ordinary practice of assigning truth values, we should surely say that the reference change takes place as soon as the new usage is stable and internalized, whatever this precisely amounts to, just as we say that the meaning of "planet" changed (more or less) instantaneously in 2005, when the IAU decided to change the definition (assuming that the rest of us in fact do defer to the IAU on this matter). But the crucial question is: can we really say this without abandoning semantic externalism? I think we can. In the rest of this section, I will explain how, and in doing so also clarify the relevant difference between slow switching cases and the kind of intentional meaning control consistent with externalism.

A semantic externalist is committed to saying that the meanings of (at least some) terms are partly determined by external matters of fact. Given a term *t*, the meaning (and thereby intension) of which is externally determined, we should separate two questions:

(1) *What kind of* external matters of fact are relevant for determining the intension of *t*, and *how*?

(2) What *are* the relevant facts pointed at, in our answer to question (1)?

For example, if we accept anything roughly like the Putnamian view of "water," the answer to (1) is: the chemical constitution of the local watery stuff matters; sharing this is necessary and sufficient for belonging in the extension of "water."[6] The answer to (2), on the other hand, is: the chemical constitution of the local watery stuff is H_2O. Something structurally similar will hold for all natural kind terms, if Putnam is to be believed.

Typically, we have very little or no control over the answers to question 2. There is little we can do about the molecular structure of the watery stuff on Earth. It is precisely this lack of control that motivates doubts about combining conceptual engineering and semantic externalism. However, this leaves open the possibility that we *may* have control over the answers to question 1: we may have control over *which* (and even *whether*) external matters of fact are relevant for determining the intension of a given term, and how such external matters of fact affect the intension. If our pretheoretical judgments regarding correct assignment of content and truth value are to be trusted, my thought experiment illustrates that we, at least in some imaginable cases, *do* have such control: the speakers in the thought experiment collectively changed the answer to question (1) to (roughly): "no external matters are relevant," thus removing the relevance of any answer to question (2), for "water."

Note also that this is not at all what happens in slow switching cases! In slow switching cases, the answer to (1) remains unchanged: what changes is the answer to question (2). The relevant changes in slow switching cases are *by hy-*

[6] It is not obvious that this Putnamian view is correct: some recent empirical evidence suggests that ordinary speakers take sharing the chemical constitution of the local watery stuff necessary, but *not sufficient* for belonging in the extension of "water" (cf. Haukioja, Nyquist & Jylkkä 2021). Such details concern, however, only the precise contents of the correct answers to (1), and do not affect the main point of this paper.

pothesis changes that are, at least ordinarily, beyond our control, and that can happen without the relevant speakers becoming aware of them. Once we notice that the answers to question (1) are just as relevant for determining the intensions of our terms, and that there is no *prima facie* reason to think we lack control over these, the tension between semantic externalism and conceptual engineering should begin to seem much less serious.

Here is another way to put the point. Semantic externalism claims that the supervenience basis of meaning includes external factors, such as the actions of other speakers, facts about underlying natures, and so on. Typically, we have little or no control over these external factors. But we may, nonetheless, have control over *what kinds of facts* are included in the supervenience basis that determines the meaning of a given term. Exactly what determines the supervenience basis for a given term is an enormously complex issue that I cannot hope to settle here, but the following rough sketch seems plausible to me, both when applied to the thought experiment above, and when considered in the abstract. The supervenience basis for a given term—which factors enter into determining its meaning—is dependent on (relatively) stable patterns of use, or perhaps stable patterns in dispositions to use, the term in question. What makes it the case that a given term has an externally determined meaning—and thereby that there exists a positive answer to question (1) for that term—is that the speakers using the term are disposed to *treat* some external facts as relevant when evaluating whether something falls under the extension of the term. If Putnam is right, the meaning of "water" is partly dependent on the fact that our local watery stuff consists of H_2O. What makes it the case that it is *this* external fact which partly determines the meaning, rather than some other external fact, or no external fact at all, is the fact that ordinary speakers (or, perhaps, expert speakers that ordinary speakers are disposed to defer to) are disposed to take information about the underlying nature of the local watery stuff as *relevant* for evaluating the correctness of the use of "water." For many other terms, such as "bachelor," we do not have similar dispositions: we would not take information about underlying properties of local

bachelors to be relevant for evaluating the correctness of using "bachelor."[7]

This kind of a view—which can be fleshed out in more systematic detail by dispositionalist theories in *meta-metasemantics*, such as (Cohnitz & Haukioja 2013) and (Johnson & Nado 2014)—suggests that answers to question (1), for terms with externalist metasemantics, are at least in principle in our control. The kind of coordinated action described by Koch *can* change the meanings of our terms even if semantic externalism is true—in fact, it can change meanings much faster than Koch himself is prepared to allow.[8] There may be all kinds of *practical* difficulties in getting people to change the ways they speak, but a systematic change in how we are disposed to speak and interpret others can change meaning, and semantic externalism does not pose a principled obstacle.

4. "This is not externalism!"

Herman Cappelen (2018) considers, and dismisses, a position much like the one I sketched in the previous section. His main target is Peter Ludlow, who argues that "it is within our control to defer to others on elements of the meaning of our words [...] and it is also within our control to be receptive to discoveries about the underlying physical structure of the things we refer to" (Ludlow 2014, 84). Cappelen replies:

---

[7] There are some who would apparently disagree (see Biggs & Dosanjh 2021), but a discussion of their view will have to wait for another occasion.

[8] A dispositionalist view can also provide an explanation of *when* meaning change occurs in slow switching cases, in terms of the relevant speakers' total dispositional states. The reason why Koch's Young-Mary, recently transported from Earth to Twin Earth and unaware of the chemical differences, refers to H_2O with her "water" is, arguably, that were she to learn of the differences, she would *retract* her application of "water" to the watery stuff on Twin Earth. The reason Old-Mary, on the other hand, refers to XYZ is that she would *not* so retract her usage. The meaning change occurs (possibly in a gradual fashion), as her dispositions to retract change. For a more detailed and systematic explanation of meaning change along these lines, see Cohnitz & Haukioja, forthcoming. For a similar account, see Nyquist 2020.

> Here is a way to understand Ludlow's position: […] what makes it the case that externalism is true is that we, in a particular conversational setting, decide that it is. According to Ludlow, if a form of externalism is true for a conversation at a time […], that is because the conversational participants […] want it to be true at that time – because they choose to defer to whatever external factors the relevant form of externalism appeals to.
>
> […]
>
> This, however, is not externalism. Externalism as I have understood it […] is not the view that conversational partners at any point in time can just decide that externalist constraints on semantics don't apply. (Cappelen 2018, 166–167)

I am not going to defend Ludlow's theory, specifically, against Cappelen's charge here (but I do believe Cappelen's characterization of Ludlow's view to be uncharitable). When it comes to the view I've sketched—which also claims that it is in a real sense within our control whether we defer to others, or are receptive to empirically discoverable factors in assigning meanings to our terms—it should be obvious that Cappelen's criticisms are off the mark. Meanings, including whether and how they are dependent on external factors, are determined by systematic and relatively stable patterns of dispositions among language users. These cannot be changed at a whim: the relevant dispositions are relatively automatic and not based on conscious deliberation. We have reason to expect that such dispositions are difficult to change. But, unlike Cappelen seems to assume, we are not faced with a choice between no control at all on the one hand, and freely chosen (meta)semantics on the other.

The interesting question is whether semantic externalism presents a *principled* obstacle for meaning control and conceptual engineering. I've argued that it doesn't. It may well be that successful conceptual engineering is hard to carry out, but that was to be expected. I hold that my thought experiment about frequent travel between Earth and Twin Earth, though no doubt fanciful in its content, nonetheless presents a case where speakers would have a widely shared practical motivation for changing the meaning of "water." Given the motivation, I think it would be realistic to expect that they

would engage, and succeed, in the kind of coordinated action required for changing the meaning. That a term has externalist metasemantics may affect *how* its meaning is to be intentionally changed, but it does not preclude *that* we can intentionally change its meaning.[9]

NTNU Trondheim

References

Biggs, Stephen, and Ranpal Dosanjh (2021). "Pervasive Externalism." In S. Biggs and H. Geirsson (eds.), *The Routledge Handbook of Linguistic Reference*. New York: Routledge, 309–323.

Burge, Tyler (1979). "Individualism and the Mental." *Midwest Studies in Philosophy* 4, 73–122.

Burge, Tyler (1988). "Individualism and Self-Knowledge." *Journal of Philosophy* 85, 649–663.

Burgess, Alexis, and Plunkett, David (2013). "Conceptual Ethics I." *Philosophy Compass* 8, 1091–1101.

Cappelen, Herman (2018). *Fixing Language*. Oxford: Oxford University Press.

Cohnitz, Daniel, and Jussi Haukioja (2013). "Meta-Externalism vs. Meta-Internalism in the Study of Reference." *Australasian Journal of Philosophy* 91, 475–500.

Cohnitz, Daniel, and Jussi Haukioja (forthcoming). *Foundations for Metasemantics*. Oxford University Press.

Devitt, Michael (1981). *Designation*. New York: Columbia University Press.

Evans, Gareth (1973). "The Causal Theory of Names." *Proceedings of the Aristotelian Society Supplementary Volume* 47, 187–225.

Haslanger, Sally (2006). "What Good are Our Intuitions?" *Proceedings of the Aristotelian Society Supplementary Volume* 80, 89–118.

Haukioja, Jussi, Mons Nyquist, and Jussi Jylkkä (2021). "Reports from Twin Earth: Both Deep Structure and Appearance Determine the Reference of Natural Kind Terms." *Mind & Language* 36, 377–403.

Johnson, Michael, and Jennifer Nado (2014). "Moderate Intuitionism: A Metasemantic Account." In A. Booth and D. Rowbottom (eds.), *Intuitions*. Oxford: Oxford University Press, 68–90.

---

[9] I am very grateful to Daniel Cohnitz, Jeske Toorman, and members of the NTNU reading group on conceptual engineering, for helpful comments on previous drafts of this paper.

Koch, Steffen (2021). "The Externalist Challenge to Conceptual Engineering." *Synthese* 198, 327–348.

Ludlow, Peter (2014). *Living Words*. Oxford: Oxford University Press.

Nyquist, Mons (2020). "On Complete Information Dispositionalism." *Philosophia* 48, 1915–1938.

Pinder, Mark (2021). "Conceptual Engineering, Metasemantic Externalism, and Speaker-Meaning." *Mind* 130, 141–163.

Putnam, Hilary (1975). "The Meaning of 'Meaning'". In *Philosophical Papers, Vol. 2: Mind, Language and Reality*. Cambridge: Cambridge University Press, 215–271.

Sawyer, Sarah (2018). "The Importance of Concepts." *Proceedings of the Aristotelian Society* 118, 127–147.

Fictional Names Revisited

PANU RAATIKAINEN

1. Introduction

Fictional names (and related thoughts) have long puzzled philosophers. Fictional entities are often used in philosophy as paradigms of something that does *not* exist, and fictional names as stock examples of names that *fail* to refer to anything. Yet we seem to be talking about *something* when we talk about, say, Sherlock Holmes. That is, sentences such as:

> Sherlock Holmes smoked a pipe
>
> Sherlock Holmes never existed

seem perfectly meaningful, and at least the latter seems true. Devitt (1981) has distinguished here two kinds of attitudes among philosophers. He calls those who insist that fictional names fail to name anything "the tough philosophers," and those who rather think that fictional names do refer to something "the tender philosophers." It seems that from Russell to Quine and beyond, "the tough philosophers" have dominated—at least in the so-called "analytic" tradition. Russell famously wrote:

> [T]o maintain that Hamlet, for example, exists in his own world, namely, in the world of Shakespeare's imagination, just as truly as (say) Napoleon existed in the ordinary world, is to say something deliberately confusing, or else confused to a degree which is scarcely credible. There is only one world, the 'real' world: Shakespeare's imagination is part of it, and the thoughts that he had in writing Hamlet are real. So are the thoughts that we have in reading the play. But *it is of the very essence of fiction that only the thoughts, feelings, etc., in Shakespeare and his readers are real, and*

> *that there is not, in addition to them, an objective Hamlet.* (Russell 1919, 169; my emphasis)

However, in the 1970s, the tide turned. In 1973, Kripke gave his famous *John Locke Lectures* in Oxford. In those lectures, Kripke defended the "tender" view. The transcript of the lectures was circulated in the philosophical community, and they influenced many philosophers. Several other philosophers, partly independently and partly influenced by Kripke, began to hold similar views. Kripke's lectures were finally published in 2013 as *Reference and Existence* (Kripke 2013). Their publication has given new currency to Kripke's particular views (see also Kripke 2011).

In those lectures, Kripke contends, at least tentatively, that fictional entities do exist as abstract objects, and fictional names do refer to such abstract entities. Several other philosophers have since then favored similar views. There is no established terminology here. I shall call this general view *The Abstract Object Theory (*AOT, in short).[1] However, we need to distinguish two different variants of the view. First, Wolterstroff (1980) has proposed a view that is somewhat Platonist: according to him, fictional entities are collections of properties and as such, *eternally* existing kinds. Second, Kripke and several others, including Searle (1975), van Inwagen (1977, 1983, 2003), Thomasson (1997, 2003), Salmon (1998), and Braun (2005), have rather held that fictional objects are *created*, and have "a time of birth." This specific view is sometimes called "Creationism." The focus in what follows is mainly on this latter view, but as "Creationism" as a label has some unhappy connotations, I prefer to talk simply about "AOT."

Advocates of AOT have utilized different analogies. Van Inwagen compares fictional entities to theoretical entities in various sciences and contends that fictional characters are theoretical entities of "criticism" or "fictional discourse." I cannot help feeling, however, that the analogy is quite weak: theoretical entities in science are typically postulated in order to explain some otherwise unexplainable observable phenomenon. Nevertheless, it is not clear what the latter would be in the case of fictional entities. In science, it is also possible

[1] "Fictional realism" is also sometimes used in the literature for this view.

that it turns out the postulated theoretical entities do not in reality exist: this happened, for example, for the postulated planet Vulcan. It not clear that the same could happen for fictional entities in the AOT framework. In practice, van Inwagen's grounds appear to reduce to the observation that we sometimes seem to quantify over fictional entities. I will set van Inwagen's analogy with theoretical entities aside, but I will return to the question of quantification later.

Kripke and Thomasson, in contrast, view fictional entities as social entities—or, at least, they defend AOT by comparing fictional entities to social entities, such as nations or laws. Kripke says:

> The fictional character can be regarded as an abstract entity which exists in virtue of the activities of human beings, in the same way that nations are abstract entities which exist in virtue of the activities of human beings and their interrelations. (Kripke 2011, 63)

> They exist in virtue of certain activities of people just as nations do. (Kripke 2013, 73–74)

Thomasson, in turn, writes:

> …the best view of what fictional characters are … is that fictional characters are abstract cultural artifacts, relevantly similar to other social and cultural entities including particular laws of state…, works of music…, and the works of literature in which fictional characters appear. (Thomasson 2003, 220)

There is one peculiar aspect in Kripke's meditations: Kripke gives quite a lot of weight in his argumentation to iterated fiction, i.e., fictions inside fictions, and related characters. He seems to think that analyzing them requires AOT. Kripke reflects on two examples. First, he considers *Hamlet* and Gonzago. Namely, in Shakespeare's play *Hamlet*, there is a play (inside the play) called *The Murder of Gonzago*. (Such a play has apparently never existed.) Inside the story, Hamlet is a real person, but Gonzago is a fictional character. According to Kripke, in our reality, "real life," Hamlet exists as an abstract object, but Gonzago apparently does not exist. Gonzago is what Kripke calls "a fictional fictional character." Second, Kripke considers Moloch, the famous pagan god whom the

Canaanites apparently worshipped. Child sacrifice and fire are often associated with Moloch. However, as Kripke points out, some scholars now contend that this is all confusion: according to them, "MLK" in ancient Hebrew just meant either "king" or "lord," or the kind of "sacrifice," and did not name a particular pagan god at all.[2] Kripke concludes that if so, there was no such pagan god as Moloch. But there really was, say, Astoreth[3] (a pagan goddess) as an abstract object (fictional entity).[4] Kripke seems to suggest that AOT is needed to make such distinctions between real and existing fictional entities, such as Hamlet and Astoreth, and characters such as Gonzago and Moloch, which do not exist in "real life."[5]

The philosophical literature on fictional names and fictional entities is now vast and has many ramifications; it would not be realistic to try to cover it comprehensively in one paper. My aims here are much more limited: I want to discuss critically especially the above-mentioned ideas of Kripke and, to some extent, the related ideas of a few other philosophers, which have not received much attention. Kripke's arguments are always to be taken seriously, and he certainly makes a number of apt observations here too. Nevertheless, I find it difficult to agree with him on this particular issue, i.e., on AOT. This paper is my attempt to spell out my reasons for that.[6]

---

[2] As Kripke notes in various footnotes, these are controversial theories, but let us assume with Kripke, for the sake of argument, that something like this is true.

[3] Kripke does not talk about Astoreth but only about Zeus; I have introduced Astoreth to make the two cases more directly comparable; apparently Astoreth was actually worshipped by the Canaanites.

[4] However, Kripke does not in the end treat Moloch (assuming that some of the critical theories Kripke mentions are correct) as a fictional fictional entity; he compares "Moloch" to failed names such as "Vulcan"—both are, according to Kripke, empty names—and says that the whole idea that there is such a legendary object is "based on a *confusion*" (Kripke 2013, 78).

[5] For example, in the Buenos Aires workshop (2013), if my memory does not fail me, Kripke seemed to take this as a major advantage of AOT, speaking in its favor. I must admit that I find the relevant passages in Kripke 2013 and Kripke 2011 puzzling.

[6] For some further complementary critical arguments against AOT or "creationism" in general, see Caplan 2004 and Brock 2010.

2. Fictional entities as social entities?

As we have noted, at least Kripke and Thomasson suggest that (real) fictional entities are more or less similar to social entities, such as *nations* (Kripke) or particular *laws* of state such as the U.S. Constitution or the Miranda laws (Thomasson), "which exist in virtue of the activities of human beings and their interrelations" (Kripke).

On the one hand, it seems evident that for a *social entity* to exist, to be real,[7] some sort of collective intention—commonly accepted rules and norms, habits, practices, and regular behavior patterns—of sufficiently many people is required. That is why they are called *social* entities. For example, I cannot myself alone create a new nation, legislate a new social norm or law, or make, say, pinecones count as currency, simply by entertaining the idea; some kind of recognition by some other people, a collective acceptance or agreement, is required. So, are fictional entities something like that?

On the other hand, Kripke and Thomasson (and some other advocates of AOT) also clearly think that a fictional entity is created and begins to exist as soon as the author writes the relevant story that first introduces the character in question, in the act of pretense. Kripke writes, for example: "On my view, to write a novel is, ordinarily, to create several fictional characters" (Kripke 2013, 72). This in no way requires any audience, any sharing with a wider community. Therefore, the fictional entity can't really be, in this picture, a social or cultural entity in any normal sense. This amounts to—if not a plain contradiction—at least a serious tension within AOT, as Kripke and Thomasson, for example, develop it. What about fictional texts that are never published?[8] How about texts that

[7] There are obviously also eliminativist views on social entities, but here only the views according to which social entities are, in some sense, real and exist, are relevant.

[8] Interestingly, Salmon writes: "Kripke believes that a fictional character does not come into existence until the final draft of the fiction is *published*" (Salmon 2011, 69, fn. 24; my emphasis; Salmon himself disagrees). I can't find anything in Kripke 2013 or Kripke 2011 that would support this; rather, they seem to support my interpretation here. But perhaps Kripke has later qualified his view this way, and Salmon has some first-hand

are never read by anyone else but the author? Furthermore, *writing* the story down cannot be essential: as Kripke (2013, 71) also notes, fictional folktales have often been passed orally from one generation to the next without them being written down. However, if neither community nor writing down is required, it seems to follow that any entity ever imagined exists (as an abstract object of imagination). But this seems excessive: it brings with it the metaphysical problem of the overpopulation of the realm of existing things.[9]

Furthermore, it seems that the latter liberal line (according to which any entity every imagined exists) would collapse Kripke's central distinction between fictional and fictional fictional entities: for surely *a* cannot imagine that *b* imagines that *P* without *a* herself imagining that *P*. For example, Shakespeare cannot imagine that an unnamed author had imagined Gonzago (and his murder) without himself imagining Gonzago. But then also Gonzago and not only Hamlet should exist as an abstract object.

Then again, Kripke wants to think that fictional entities do *not* exist "automatically," but it is an empirical question whether a certain fictional entity exists (Kripke 2013, 71):

> Was there a fictional or legendary character who married his grandmother? ... If there was, this will be true in virtue of appropriate works of fiction or legend having been written, or at least told orally, or something of the kind. If there is such a fictional work, then there is such a fictional character. (Kripke 2013, 71)

> The question of their [fictional characters'] existence is a question about the actual world. It depends on whether certain works have actually been written, certain stories in fiction have actually been told. (Kripke 2011, 63)

---

knowledge of that—Salmon refers in another footnote to Kripke's seminars he attended in 1981 and 1983. Be that as it may, there are critical questions also in that case: What if nobody reads the published fiction? How about orally transmitted folktales or widely circulated manuscripts that never get published? Clearly being published cannot be a plausible demarcation line here.

[9] I owe the key observations of this paragraph to Jenni Tyynelä.

Here it seems that the existence of a fictional entity depends on the existence of the relevant *fictional work*—where the latter seems for Kripke here to require something more substantial than just someone momentarily imagining that entity. For example, apparently the made-up play *The Murder of Gonzago* does not exist by Kripke's standards, and it would seem to follow that Gonzago does not therefore exist as an abstract object in the real world. But this does not cohere with the liberal conclusion we ended up with just a moment ago. I repeat: there is a serious tension here.

3. Fiction inside fiction

Let us next reflect on iterated fiction and related issues in more detail. Do such cases support AOT? To begin with, Kripke's example of Moloch seems to be a bit off-topic. It is not really a case of a fictional work and its content; it essentially turns to the factual historical question of whether certain people in a certain place and time really believed in and worshipped such-and-such a pagan god in such-and-such a way. It may be a false historical hypothesis that there was such-and-such a religion or cult in the land of Canaan around 1200–800 BC, that the Canaanites worshipped at the time such-and-such a pagan god, etc. But I fail to see why any of this would require us to think that, in contrast, say, Astoreth, exists or existed (apparently, she was really worshipped by the Canaanites at that time). I think we can analyze quite easily the sentence

> Canaanites, around 1200–800 BC, worshipped a god of fire, called "Moloch" (or something like that), essentially by sacrificing children to him by burning them

as false (if the above-mentioned hypothesis is correct) without being required to conclude that Moloch does (or did) not exist, but Zeus and Astoreth, in contrast, do (or did) exist. These are questions about certain specific groups of people, in a specific time period, and whether they held such-and-such beliefs, whether they practiced such-and-such religions, worshipped such-and-such gods in such-and-such a way, etc. It is neither natural nor very helpful to interpret this as a question of whether this or that god exists (or existed) or not.

Anyway, these are factual questions of beliefs, and not questions about imagination and fiction.

And why, if AOT is correct, instead of saying that Moloch never existed (as Kripke suggests), do we not say that he/it did not exist around 1200–800 BC, but was created later and has existed? After all, the idea of Moloch has been later widely shared in the Jewish and Christian world. Compare this to the following scenario: Imagine that it turned out that *Hamlet* was *not* really written by Shakespeare around 1600; instead, there had been an ingenious hoax and the work had been written only in the 19th century. Should we conclude, according to AOT, that Hamlet never existed? Or just that Hamlet was created later and has existed for a shorter time than we had assumed?[10] In sum, it is quite unclear whether and how the Moloch case supports AOT (if that ever was Kripke's idea).

Kripke's alleged conclusion that Gonzago, in contrast to Hamlet, does not exist may feel intuitively appealing, because *Hamlet* (the play) leaves the fictional fictional story of Gonzago so sparse, superficial, and incomplete: we are told very little about Gonzago in *Hamlet*, and the character is left highly unspecified. But I think that our *prima facie* intuitions may vary depending on the vividness or the specificity of the fictional fictional story.

For example, let us rather consider *The Taming of the Shrew*, another famous play by Shakespeare. In it, the frame story is quite short and unspecified. Its main character is a drunk tinker named Christopher Sly. When he wakes up, he is tricked, as a prank, into believing that he is actually a nobleman. A play is then performed for Sly that includes as characters two daughters of the rich lord Baptista Minola and several of their suitors. It is this play inside the play that is the main plot developed in detail, and which is best-known, having vivid characters such as the wild daughter Katherina and her harsh suitor Petruchio. I think we are much less inclined to conclude that only Sly really exists as an abstract fictional entity but that Katherina and Petruchio do not. But this is

[10] In Kripke 2011, in a footnote (29) added presumably somewhat later (than the original 1973 talk), Kripke reflects briefly on what is seemingly the same point.

what Kripke's view would presumably require us to conclude if we were to follow it consistently.

"The Grand Inquisitor" in *The Brothers Karamazov* by Dostoevsky is one of the best-known passages in literature, and the Inquisitor himself a very well-known character. However, the tale is a fiction inside fiction, a story told by Ivan, one of the fictional brothers. Again, I think it is not obvious—if we were ever to accept fictional entities as abstract objects to our ontology—that the famous Inquisitor does not exist, but only the brothers and their father do. In *Winnie-the-Pooh* by A. A. Milne, the brief frame story includes only Christopher Robin and the narrator (apparently his father).[11] The narrator then tells stories about the adventures of Winnie-the-Pooh and other familiar characters to Christopher Robin. Again, if we accept fictional entities as existing abstract objects at all, it is far from obvious that we should only accept Christopher Robin and the narrator as such, but not the famous Winnie-the-Pooh, Piglet, etc. Consider also *One Thousand and One Nights*: In its frame story, Shahrazad tells tales to her husband Sultan Shahryar over many nights. The best-known characters, such as Aladdin and Sinbad, occur in these tales and are only "fictional fictional characters" (in Kripke's sense).

I contend that if we accept fictional entities into our ontology at all, we should certainly include, for example, the Inquisitor, Winnie-the-Pooh, and Aladdin, and not only the characters of the frame stories. The former world-famous characters are cultural entities if anything is, even if they are only "fictional fictional entities" by Kripke's standards. But if so, AOT cannot then be used to make the distinction between fictional entities (which, according to AOT, exist) and fictional fictional entities (which allegedly do not exist), as Kripke seems to suggest. However, then it cannot be used as an argument in favor of AOT that it enables a line to be drawn between them in the first place.

Further, if it supports AOT that it allows us to distinguish fictional characters from fictional fictional characters, what about extra iterations? For example, in *One Thousand and One*

[11] For simplicity, I shall ignore the historical fact that they were apparently modelled after Milne and his son, and treat them as purely fictional characters here.

Nights, one of Shahrazad's (level 1) stories is the tale of the Fisherman and the Jinni (level 2). In that story, the fisherman then tells the Jinni the tale of the Vizier and the Sage Duban (level 3). And in this story, King Yunan in turn tells the Vizier the tale of the Husband and the Parrot (level 4). Kripke's AOT, which seems to classify both the fisherman (level 2), the Vizier (level 3), and the husband (level 4) as non-existent, cannot as such distinguish these levels from each other.[12] Clearly we need some other, simpler, and more fundamental way to keep track of the different levels of fictional stories. But presumably we can then also distinguish levels 1 and 2 without having to assume that the characters of level 1 exist as abstract objects but the characters of level 2 do not exist.

In sum, postulating fictional entities as abstract objects cannot be supported—if that was Kripke's intended argument—with the help of the differences between fictional tales and tales inside such tales. In any case, Kripke seems to put too much weight on this distinction. If an author creates, in whatever sense, first-level fictional entities, he or she similarly creates the second-level entities which are fictional fictional. There does not seem to be any principled metaphysical difference between them: either both exist, or neither of them do.

4. Quantification over fictional entities

The fact remains that we often seem to quantify over fictional entities. Does this mean that we are thereby ontologically committing ourselves to the existence of fictional entities? Kripke and especially van Inwagen suggest that we do, and that this supports AOT. I contend that this issue requires closer examination. (However, I must necessarily be rather brief and selective here.)

[12] To be sure, one could say that *inside* the frame story, Shahrazad and Shahryar are real persons but the Jinni, for example, is an abstract fictional entity, and the Vizier, as a fictional fictional entity, does not exist. Similarly, *inside* the tale of the Fisherman and the Jinni, the Jinni is a real entity but the Vizier is an abstract fictional entity, and the Husband does not exist. But it is quite unclear what would be achieved with such a complicated way of talking, or whether it is in any way necessarily required.

To begin with, as Kripke already clearly notes, there are two importantly different contexts of quantification here. First, there is quantification within fiction:

> In the fictional story *S*, there is a detective such that...

But second, there are statements *about* fiction, i.e., uses outside the scope of the imagining. For example, from the apparent fact that Sherlock Holmes was created by Arthur Conan Doyle, one could conclude:

> There is an x such that x is Sherlock Holmes and x was created by Conan Doyle.

Whereas the former may be quite harmless, as the apparent quantification occurs inside the fictional story and is part of the pretense in general, the latter cannot be that easily swept under the rug. These are the cases that, according to some philosophers, really commit us ontologically to fictional entities. In the rest of this section, I shall argue that this is not necessarily the case.

There have obviously been attempts to avoid this conclusion. Yagisawa (2001) has suggested that perhaps quantification in such cases could be interpreted as substitutional quantification and not as standard objectual quantification. I am not, however, convinced that the strategy could work in general. One problem is that fictional stories frequently include unnamed characters. Then again, Priest (2005) and Crane (2013), for example, have proposed the radical view that we should in general give up the association between quantification and ontological commitment.[13] I am not entirely unsympathetic toward such proposals; there is much to be said on their behalf. However, at least in the case of fiction in particular, a less radical approach seems sufficient. Namely, clearly fiction is closely related to imagination. They both result in *intensional contexts*:

> In the fictional story *S*, *p*

[13] In *The John Locke Lectures* (Kripke 2013), Kripke shows some sympathy towards the idea that existence would be treated as a predicate, and thus separated from quantification. It is not clear to me how this harmonizes with the idea that quantification over fictional entities speaks in favor of the existence of fictional entities.

a imagines that *p*

The latter is a kind of propositional attitude, the intensional logic of which has been pursued for many decades. It is well-known that in intensional contexts quantification is not necessarily ontologically committing. This is the case already in the standard alethic modal logic (i.e., the logic of necessity and possibility): there are strong pressures either to adopt free logic or to treat objects in the domain of quantification as merely possible objects, not necessarily as objects that actually exist (see, e.g., Garson 1984).

In the case of propositional attitudes, things are even more complicated. Intensional logics and possible world semantics for them have been developed especially by Hintikka and his followers. Hintikka (1962) has famously presented logic for *knowledge* and *belief*. However, it was his logic of *perception* (see Hintikka 1969, 1975; cf. Niiniluoto 1979, 1982; Saarinen 1987), in connection to which Hintikka developed certain insights on quantification, that we will discuss shortly. It was left to Niiniluoto (1983, 1986) to develop a similar logic for *imagination* as a propositional attitude.

As to perception and hallucination, David Lewis once wrote:

> What do we see when we see what isn't there?
>
> Macbeth the hallucinator sees a dagger. There is no dagger there to be seen ... There is no reason to think that our world contains any such thing. But the lack of a dagger makes it mysterious how we can describe Macbeth's state, as we do, by means of predicates applying to the dagger he seems to see ...
>
> The case of the missing dagger has been solved by inspector Hintikka. I accept his solution... (Lewis 1983, 3)

I contend, following Niiniluoto (1983, 1986), that a related solution can be given for the mystery of objects of imagination and fiction. So, let us review how apparently the most well-developed and sophisticated account of imagination and

quantification available treats them.[14] Along familiar lines, we can first stipulate:

> *a* imagines that *p* = in all possible worlds[15] compatible with what *a* imagines it is the case that *p*.

Let us write I_a *p* for "*a* imagines that *p*." But how is an individual identified in different possible worlds? Intuitively, a "world-line" is a line which connects, somewhat like connecting dots, one and the same individual from different worlds. Formally, a "world-line" is a function from worlds to individuals. In the Hintikka-style logic of propositional attitudes, quantified variables range over these world-lines.

A crucial observation here is the following: the sentence "*a* imagines that *b* is *F*," for example, cannot always be adequately formalized in the simple form I_a *F*(*b*), but in order to distinguish different ways in which *b* is "presented" to *a*, or *a*'s act of imagining is "directed" to *b*, we need (following Hintikka's logic of perception) two different kinds of quantifiers:

> (*Ex*) – physical quantifier (grounded on spatio-temporal or causal continuity)
>
> (∃*x*) – perspectival quantifier (grounded on the role in the context)

Consider, for example, the sentence:

> Michael imagines that David is dancing with a blond woman.

David here is a well-defined "physical" individual; and that can be expressed with a physical quantifier:

---

[14] I am aware of certain more recent, alternative approaches to the logic of imagination due to Wansing 2017 and Berto 2017. However, their focus is more on the voluntary nature of typical imagination, the well-known failure of logical closure, and such. In any case, they both restrict their attention to propositional logic. Consequently, whatever their virtues, they cannot really illuminate better the behavior of quantification in the context of imagination.

[15] "Possible world" must here be understood in a liberal sense of "possible": many of them may *not* be metaphysically possible in the standard Kripkean sense.

$(Ex)\ \mathrm{I_M}\ [x = \text{David} \wedge x \text{ dances with a blond woman}]$.

However, "the blond woman" is (let us assume) non-specific. She could be, in different worlds (it is compatible with what Michael imagines that she would be), for example, Marilyn Monroe, Dolly Parton, or Debbie Harry, but also (to consider some purely fictional possibilities) Beatrix aka "The Bride" (in *Kill Bill*), Cathrine Tramell (in *Basic Instinct*), Pussy Galore (in *Goldfinger*), etc. They play, in different possible worlds, the same relevant role in what Michael imagines. The perspectival world-line picks out different blond women from different worlds: the blond woman that plays the same role in Michael's field of imagination as the dancing partner of David in that world. This can be expressed with a perspectival quantifier:

$(\exists y)\ \mathrm{I_M}\ [y \text{ is a blond woman} \wedge \text{David is dancing with } y]$.

And putting the above two together:

$(Ex)(\exists y)\ \mathrm{I_M}\ [x = \text{David} \wedge y \text{ is a blond woman} \wedge x \text{ is dancing with } y]$.

The following two:

$(\exists x)\ \mathrm{I_a}\ (x = b)$ – *a* imagines of *b* something;

and

$(\exists x)\ \mathrm{I_a}\ F(x)$ – *a* imagines an *F*;

do not entail that *b*, or anything that is *F*, exists or is real; they can cover both existing and non-existing entities. In contrast, in the following three cases, even if the perspectival quantifier is used, the object of imagination, *b*, must actually exist:

$(\exists x)\ [x = b \wedge \mathrm{I_a}\ (\exists y)(x = y)]$ – *a* imagines something about *b*.

$(\exists x)\ [x = b \wedge \mathrm{I_a}\ (x = c)]$ – *a* imagines *b* as *c*.

$(\exists x)\ [x = b \wedge \mathrm{I_a}\ F(x)]$ – *a* imagines of *b* that she is an *F*.

This is because "*b*" is outside the scope of the operator "$\mathrm{I_a}$," and its occurrence is transparent. It is also possible to state explicitly, with this formalism, that the object of imagination does not exist:

$(\exists x)\ \mathrm{I_a}\ [x = b \wedge F(x)] \wedge \neg\ (\exists x)(Ey)[x = y \wedge \mathrm{I_a}\ (x = b \wedge F(x))]$.

Here the imagined "something" does not exist, not even as an abstract object, in the actual world, nor as a possible object. It is not an object in any particular world, but only a "world-line" that does not continue to the actual world. In Hintikka's words, such objects are "neither here nor there."

Obviously, much more could be said concerning this issue, but that must be left for another occasion. Here I content myself with noting that at least in the context of one of the most well-developed theories in this area, quantifying in the contexts of imagination and in fictional contexts does not force us to assume that the relevant objects of imagination must exist.

As we have noted above, Kripke and Thomasson have compared fictional entities to created social and cultural entities like nations and works of art. Niiniluoto, for his part, has argued that as to the issue of their reality, we should not conflate fictional works of art and fictional entities within the former. He has appealed here to Peirce's "scholastic" criterion of reality, according to which those things are *real* "whose characters are independent of what anybody may think them to be" (CP 5.311, 5.405). Peirce himself applied this definition in 1878 to distinguish reality and fiction (as opposites), e.g., the fact of my dreaming may be real while the things dreamt are not (CP 5.405). Accordingly, Niiniluoto (1984) has contended that when he imagines a pink elephant, his mental state is real, but the elephant is fictional, since it has only such characters that my thought impresses upon it. Later, he has extended the idea to our very topic. Niiniluoto (2006, 2011) argues that, e.g., Tolstoy's novel *Anna Karenina*—the work of art as a cultural entity—is real. However, the properties of the fictional entity Anna Karenina include only those implied by the novel. Consequently, the latter is not real. I am inclined to agree with Niiniluoto here.

5. Are fictional names ambiguous?

We do enlighten our children with statements such as:

> The bogeyman does not exist.

And such statements are, in all reason, true. AOT must somehow accommodate such obvious facts that it seems to contradict. For such reasons, at least Kripke contends explicit-

ly that fictional names are in fact *ambiguous* (Kripke 2013, 149):[16] for example, "Sherlock Holmes" in "There is an x such that x is Sherlock Holmes and x was created by Conan Doyle" and in "Sherlock Holmes never existed" has, in some sense, different meanings, and consequently both sentences can be true.[17]

It is easy to feel that this is a bit *ad hoc*, and we should in my view prefer, if other reasons do not force us to that conclusion, a theory which does not require that fictional names are ambiguous. This general idea is nicely captured in the following maxim referred to by Putnam: "Differences of meaning are not to be postulated without necessity." According to Putnam, Ziff calls this "Occam's eraser" (see Putnam, 1965, 130). Somewhat ironically, in the very same *John Locke Lectures*, Kripke himself mentions the maxim which says that "we are not to postulate ambiguities where they are not needed" (Kripke 2013, 125). Elsewhere, he even writes:

> It is very much the lazy man's approach in philosophy to posit ambiguities when in trouble. If we face a putative counterexample to our favorite philosophical thesis, it is always open to us to protest that some key term is being used in a special sense, different from its use in the thesis. We may be right, but the ease of the move should counsel a policy of caution: Do not posit an ambiguity unless you are really forced to, unless there are really compelling theoretical or intuitive grounds to suppose that an ambiguity really is present. (Kripke 1977, 268)

I'd like to suggest that we should follow this policy also in the case of fictional names.[18]

---

[16] Salmon (2011), on the other hand, explicitly denies that fictional names are in this way ambiguous, even if he advocates AOT.

[17] This example is, though, my own construction. Kripke (2013, 149) is talking about "Hamlet", in "Hamlet does not exist" and "Hamlet is only a fictional character."

[18] Salmon (2011), though he advocates AOT, is in complete agreement with me here.

6. Conclusion

AOT contends that fictional entities are real, existing, abstract social or cultural entities. But the other idea of AOT that the author alone creates the entity does not harmonize well with this. AOT threatens to collapse to the excessive view that any entity ever subjectively imagined even by one subject exists. Kripke's apparent suggestion that AOT makes it possible to distinguish fictional and fictional fictional entities turns out to be, on closer scrutiny, quite unclear. Surely, we need to somehow distinguish different levels in fiction, but AOT does not provide a working tool for it. Finally, quantifying in intensional contexts, such as a fictional context, arguably does not entail existence and ontological commitments.

As we have found the positive arguments in favor of AOT far from conclusive, and it apparently requires us to postulate the ambiguity of the described sort largely just to save the theory, this also speaks in favor of the "tough" view: fictional entities do not exist, just like common sense suggests, and there is no need to postulate ambiguity in the case of fictional names. Santa Claus just does not exist.

Acknowledgements

This paper is dedicated to the memory of my late wife, Jenni Tyynelä. It was she who was mainly working on this theme. Some insights in the paper are solely due to her, and the ideas which are mine emerged during our lively discussions about the topic. We were planning a joint paper about all of this, but her death came first. Still, I am solely responsible of this paper as it is.

Earlier versions of this paper have been presented at the *Philosophy of Linguistics and Language* workshop at the Inter-University Centre, Dubrovnik, in September 2018, and at the *Nordic Network in Metaphysics* seminar, in December 2021. I am grateful to the audiences for comments. I would also like to thank Michael Devitt and Tim Crane for stimulating discussions about the topic, and Tim Crane and Ilkka Niiniluoto for useful comments on an earlier draft of this paper.

Tampere University

References

Berto, Francesco (2017). "Impossible Worlds and the Logic of Imagination." *Erkenntnis* 82, 1277–1297.

Braun, David (2005). "Empty Names, Fictional Names, Mythical Names." *Noûs* 39, 596–631.

Brock, Stuart (2010). "The Creationist Fiction: The Case against Creationism about Fictional Characters." *Philosophical Review* 119, 337–364.

Caplan, Ben (2004). "Creatures of Fiction, Myth, and Imagination." *American Philosophical Quarterly* 41, 331–37.

Crane, Tim (2013). *The Objects of Thought.* Oxford: Oxford University Press.

Devitt, Michael. (1981). *Designation*. New York: Columbia University Press.

Garson, James W. (1984). "Quantification in Modal Logic." In D. Gabbay and F. Guenther (eds.), *Handbook of Philosophical Logic, Volume II: Extensions of Classical Logic*, Dordrecht: D. Reidel, 249–307.

Hintikka, Jaakko (1962). *Knowledge and Belief.* Ithaca: Cornell University Press.

Hintikka, Jaakko (1969). *Models for Modalities.* Dordrecht: D. Reidel.

Hintikka, Jaakko (1975). *The Intentions of Intentionality.* Dordrecht: D. Reidel.

Kripke, Saul (1977). "Speaker's Reference and Semantic Reference." *Midwest Studies in Philosophy* 2, 255–276.

Kripke, Saul (2011). "Vacuous Names and Fictional Entities." (A talk given at the University of Connecticut in March of 1973.) In S. Kripke, *Philosophical Troubles: Collected Papers Vol 1.* New York: Oxford University Press, 2011, 52–72.

Kripke, Saul (2013). *Reference and Existence: The John Locke Lectures* (originally delivered at Oxford University, Oct. 30 – Dec. 4, 1973). New York: Oxford University Press.

Lewis, David (1983). "Individuation by Acquaintance and by Stipulation." *Philosophical Review* 92, 3–32.

Niiniluoto, Ilkka (1979). "Knowing that One Sees." In E. Saarinen et al. (eds.), *Essays in Honour of Jaakko Hintikka*. Dordercht: D. Reidel, 249–282.

Niiniluoto, Ilkka (1982). "Remarks on the Logic of Perception." In I. Niiniluoto and E. Saarinen (eds.), *Intensional Logic: Theory and Applications*. Helsinki: The Philosophical Society of Finland, 116–129.

Niiniluoto, Ilkka (1983). "On the Logic of Imagination." In I. Patoluoto et al. (eds.), *Vexing Questions.* Reports from the Department of Philosophy, University of Helsinki, N:o 3, 23–28.

Niiniluoto, Ilkka (1984). "Realism, Worldmaking, and the Social Sciences." In I. Niiniluoto, *Is Science Progressive?* Dordercht: D. Reidel, 211–225.

Niiniluoto, Ilkka (1986). "Imagination and Fiction." *Journal of Semantics* 4, 209–222.

Niiniluoto, Ilkka (2006). "World 3: A Critical Defence." In I. Jarvie, K. Milford, and D. Miller (eds.), *Karl Popper: A Centenary Assessment vol. II. Metaphysics and Epistemology*. Aldershot: Ashgate, 59–69.

Niiniluoto, Ilkka (2011). "Virtual Worlds, Fiction, and Reality." *Discusiones Filosóficas* 12(19), 13–28.

Peirce, Charles Sanders (1931–35). *Collected Papers* (ed. by C. Hartshorne and P. Weiss), vols. 1–6. Cambridge, MA: Harvard University Press.

Priest, Graham (2005). *Towards Non-Being: The Logic and Metaphysics of Intentionality*. Oxford: Oxford University Press.

Putnam, Hilary (1965). "How Not to Talk about Meaning." In R. S. Cohen and M. R. Wartofsky (eds.), *Boston Studies in the Philosophy of Science*, Vol. 2. New York: Humanities Press, 205–222. Reprinted in H. Putnam, *Mind, Language and Reality*. Philosophical Papers: Volume 2. Cambridge: Cambridge University Press, 117–131. (Page references are to the reprint.)

Russell, Bertrand (1919). *Introduction to Mathematical Philosophy*. London: George Allen & Unwin, Ltd.

Saarinen, Esa (1987). "Hintikka on Quantifying in and on Trans-World Identity." In Radu J. Bodan (ed.), *Jaakko Hintikka.* Profiles 8. Dordrecht: D. Reidel, 91–122.

Salmon, Nathan (1998). "Nonexistence." *Noûs* 32, 277–319.

Salmon, Nathan (2002). "Mythical Objects." In J. K. Campbell et al. (eds.), *Meaning and Truth: Investigations in Philosophical Semantics.* New York: Seven Bridges, 105–23.

Salmon, Nathan (2011). "Fiction, Myth, and Reality." In A. Berger (ed.), *Saul Kripke*. Cambridge: Cambridge University Press, 49–77.

Searle, John R. (1975). "The Logical Status of Fictional Discourse." *New Literary History* 6, 319–32.

Thomasson, Amie (1999). *Fiction and Metaphysics.* Cambridge: Cambridge University Press.

Thomasson, Amie (2003). "Speaking of Fictional Characters." *Dialectica* 57, 205–223.

Van Inwagen, Peter (1977). "Creatures of Fiction." *American Philosophical Quarterly* 14, 299–308.

Van Inwagen, Peter (1983). "Fiction and Metaphysics." *Philosophy and Literature* 7, 67–77.

Van Inwagen, Peter (2003). "Existence, Ontological Commitment, and Fictional Entities." In M. J. Loux and D. W. Zimmerman (eds.), *The Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 131–57.

Wansing, Heinrich (2017). "Remarks on the Logic of Imagination. A Step towards Understanding Doxastic Control through Imagination." *Synthese* 194, 2843–2861.

Wolterstorff, Nicholas (1980). *Works and Worlds of Art*. Oxford: Clarendon Press.

Yagisawa, Takashi (2001). "Against Creationism in Fiction." *Philosophical Perspectives* 15, 153–172.

Indeterminism about Discourse Domains

TEEMU TAURIAINEN

1. Introduction

It is an uncontroversial assumption that our thoughts and speech fall within categories according to their topic or subject matter. Pre-theoretically, we distinguish between discourse about weather, politics, and interpersonal relationships, and we understand that there is something distinctive about these topics. In more formal contexts, schools offer classes on physics and mathematics, and similar distinctions are deployed in our scientific institutions, where boundaries are drawn between domains of inquiry, like physics and philosophy, and their subdomains, like ethics and aesthetics. Further, various philosophical theories rely on there being robust boundaries between discursive contents. For instance, some ethical expressivists argue that while sentences from the domain of physics are susceptible to claims about truth and falsity owing to their descriptive nature, the domains of ethics and aesthetics are non-truth-apt as they encompass primarily non-descriptive or expressive content.

However, such an argument relies on there being a robust distinction between the discourse domains of physics or descriptive discourse and ethics, aesthetics, or expressive discourse. As another example, a fact-based correspondence theorist who is a mathematical fictionalist might argue that assertions belonging to the mathematical domain are non-truth-apt as they are insusceptible to the preferred correspondence criterion for truth, assuming that there are no facts with which mathematical statements can correspond. Again, such an argument relies on there being a robust distinction between the domains of mathematics or fictional discourse and factual discourse. Finally, in more recent literature, alethic pluralists of various sorts explicitly rely on discourse

domains as an explanatory resource to support their core claim about the variability of truth across domains: "Domains are a crucial component of the theoretical framework of pluralism, as reflected by the fact that the core pluralist thesis is that the nature of truth varies across domains" (Pedersen, Wyatt, & Kellen 2018, 6).[1] Interestingly, domains have not been studied to the extent that one would expect in the current truth pluralist literature: "Despite the central role that domains play within the standard pluralist framework not much systematic work has been done on their nature" (Kim & Pedersen 2018, 111). In short, for some pluralists, sentences from distinct topically individuated domains, like physics and aesthetics, get to be true in different ways by possessing the operant truth-determining property, like correspondence or coherence for their domain.

Surprisingly, despite the widespread relevance of discourse domains for philosophical theories of various sorts, alethic theorists have said relatively little about their nature in current debates. One reason for this is that the project of defining discourse domains is similar to the challenging task of providing a philosophically tenable account of subject matters or content kinds. There are many ways to draw boundaries between discursive contents, and determining which divisions are fundamental or should be prioritized is a controversial matter. Further, as problems with defining subject matters and the domains of sentences falling within them concern a range of philosophical theories, this eases the pressure for any particular theorist to touch on this topic. Finally, the project of defining discourse domains bears an intimate connection to the notoriously challenging task of defining truth-aptness.[2] For instance, insofar as the traditional monist accounts make positive claims about the nature of truth via reference to truth-determining properties, like correspondence and coherence, such an argument involves demarcating

[1] The pluralist thesis is intuitively appealing, for it is a reasonable assumption that different kinds of sentences can be true in different ways independent of how their kinds are defined or what the specific ways of being true are.

[2] According to one view, by being maximally permissive with truth-aptness, the problem of demarcating truth-apt and non-truth-apt domains dissolves. Such a case can be made in support of the deflationary position.

sentences to truth-apt and non-truth-apt domains according to their susceptibility to the preferred criterion for truth.[3] Usually, such arguments proceed as follows: for example, a neo-classical correspondence theorist will argue that domains like physics or discourse about extensional states of affairs are truth-apt, whereas discourse about abstract entities or projected properties is not, and vice versa for coherence theorists. Hence, the traditional monist accounts also rely on there being robust boundaries between kinds of sentences, which raises a question about how such kinds ought to be demarcated.

Motivated by the current lack of research on discourse domains especially in the alethiological literature, the primary goal of this paper is to participate in the discussion on the preferred method of defining discourse domains for them to provide the sought-after explanatory utility of drawing robust boundaries between truth-apt kinds of content or content types. Based on this, central themes of discussion are the theoretical desiderata of domains to provide explanatory utility for the monists to argue for the difference between truth-apt and non-truth-apt domains, and how pluralists can explain the variability of truth across topically individuated domains, like physics and aesthetics.

The concluding argument is that insofar as domains are understood as classes of sentences that are individuated by topical subject matters, the inevitable temporal development of our topical categories and the existence of so-called mixed content compromise our ability to definitively account for the domain membership of all truth-apt contents. This creates

[3] While one might counter such an argument by pledging allegiance to some variant of deflationary theory of truth that can accommodate the truth and falsity of all syntactically proper sentences that can be supplemented to the preferred deflationary schema, the problem with the deflationary approach is that it renders either the concept of truth (conceptual deflationism) or the property of being true (metaphysical deflationism) insubstantive and unexplanatory, impeding us from utilizing truth to define other concepts, like knowledge, meaning, or validity, or understanding societally and theoretically important phenomena, like what is a general goal of inquiry that binds all the vastly different scientific disciplines or what is, in general, correct to believe and assert in epistemically relevant discourse.

confusion among alethic theorists of various sorts about the domain membership of some truth-apt sentences, subsequently generating definitional issues of various sorts. Based on this, alethic theorists specifically should seriously consider indeterminism about the extensions of fundamental domains. According to this view, while topical domains can be defined in general as relatively well-individuated classes of sentences, they are susceptible to inherent indeterminacies that leave even the more prominent accounts of domains confused on the domain membership of *some* sentences. This argument is relevant for all theorists who rely on there being robust boundaries between discursive contents.

2. Disambiguating discourse domains

Under any natural language *L*, such as English, one can find classes of linguistic objects, like words or sentences, which are individuated on the basis of a factor like topic or subject matter. In this paper, topics and subject matters are understood synonymously as semantic categories under which one finds concepts or sentences governed by the respective subject matter. The sentence "π is 3.141" is a mathematical sentence because it composes of mathematical content, and the sentence "the earth is moving" is a sentence of physics because it composes of content that is relevant for physical inquiry. This aligns with how some truth pluralists understand the nature of subject matters:

> Domains are sets of propositions individuated by their subject matter. […] ⟨2 + 3 = 5⟩, ⟨Mt. Everest is extended in space⟩, and ⟨Bob's drunk driving is illegal⟩ belong to different domains. Why? Because they concern different subject matters or are about different kinds of states of affairs. (Kim & Pedersen 2018, 112)

For the sake of clarifying the exposition, we treat the expressions falling under subject matters as *atomic sentences* of the form "*a* is *F*" ("snow is white") that consist of a singular term "*a*" (snow) designating a range of objects and a predicate "is *F*" (is white) that attributes a property to the objects designated. Thus, in the context of this paper, the discussion on the nature of discourse domains is constrained to classes of atom-

ic sentences individuated by a topical subject matter.[4] Note that as atomic sentences are generally taken syntactically as the most basic types of assertions, demonstrating problems with domains for such sentences also scales to more complex expressions, an obvious example being compounds of atomics. Further, treating atomic sentences as the constitutive contents of domains is compatible with them being interpretations of atomic sentences or atomic propositions.

Regarding the nature of subject matters, it is worth emphasizing that in the context of this paper, they are understood as topically rather than ontologically individuated categories. There are several reasons for this. Initially, the gate is open for arguing that under any natural language, one can form domains of sentences according to their representations of different aspects of the world, like ontologically distinct types of objects and properties. However, my contention is that subject matters are ordinarily taken as topical rather than ontological categories in both mundane and more formal discourses. Pre-theoretically, we regard subject matters as topical categories, and nothing prevents us from thinking that under such topics, one finds sentences about ontologically distinct aspects of the world. Similarly, in more formal discourse, we divide scientific disciplines into domains of inquiry, like physics and aesthetics, with no rules for what types of objects and properties are relevant for each domain. While we will defend this view further in the following sections, for now, it suffices to note that treating subject matters as primarily topical categories aligns with how some contemporary theorists of truth understand the nature of subject matters. Instead of discourse domains, Lynch (2009, 77–79; 133) discusses domains of inquiry, like physics and ethics. Furthermore, Wyatt argues as follows:

> There is, for instance, distinctively mathematical subject matter: sets, numbers, the successor function, and so on. There is also a class of propositions that are mathematical in kind: ⟨the null set has zero members⟩, ⟨the successor of 1 is 2⟩, and so on. These propositions are mathematical propositions because they are

[4] There currently exists no general theory of sentential topics or subject matters in contemporary literature on the philosophy of language.

> composed of mathematical concepts, i.e., concepts about the subject matter mathematics. (2013, 230)

As Wyatt adequately notes, it is a reasonable assumption that sentences belong to domains by composing of kinds of concepts, where these kinds are understood in terms of topical subject matters. In this sense, there are topically defined subject matters, like physics and aesthetics, which govern a range of concepts about or falling under the relevant subject matter. Further, sentences assign as members of topically individuated domains by composing of the aligning concepts. Consequently, we understand discourse domains as individuated classes of (atomic) sentences that belong to their respective domains by instantiating kinds of concepts, where such kinds are understood on the grounds of topical subject matters. From here, we proceed to discuss the theoretical desiderata of domains for them to provide the sought-after explanatory utility of demarcating different kinds of contents reliably to distinguish between truth-apt and non-truth-apt sentences, or sentences that are susceptible to being true in different ways.

3. Theoretical desiderata of discourse domains

There are two desiderata that domains ought to fulfill for them to provide precise boundaries for demarcating content kinds. These are *unambiguous* identities and *determinate* rules for membership. By fulfilling such criteria, domains would stand as well-individuated classes of sentences with determinate (yet potentially infinite) extensions. Note that these requirements bear an intimate connection to one another. Without unambiguous identities, it becomes difficult to define domains as classes of sentences with determinate extensions. Further, without determinate extensions, particular sentences can have confused domain membership, or they can count as members of multiple domains in an indeterminate manner, creating subsequent confusion about the identities of the respective domains.

If the aforementioned criteria are met, then domains map robust distinctions between content kinds that theorists of various sorts can utilize as a theoretical resource to explain, for example, that some domains are susceptible to truth-

aptness while others are not, or that some domains are susceptible to being true in one way rather than another.

However, one might contend that the requirement of domains as unambiguous classes of sentences with determinate rules for membership is too restrictive. One reason is the reasonable assumption that the subject matters of the constitutive concepts of sentences might be mixed or confused, or that sentences can otherwise address multiple subject matters at once and hence count as members of multiple domains with no scaling answer to which of these domains ought to be treated as primary from the perspective of their truth-aptness or way of being true. For example, nothing prevents an aesthetical theory from referring to facts pertaining to the domain of natural sciences, and similarly, though unlikely, nothing in principle prevents the deployment of aesthetical concepts as explanatory resources under one's theory about the physical properties of extensional objects.[5] However, such an argument is blind to the distinction of how we still individuate the respective sentences themselves as aesthetical and physical. In this sense, while nothing blocks one from distinguishing so-called mixed domains including contents that address multiple subject matters, we individuate the subject matter of the constitutive contents themselves on at least seemingly unambiguous grounds. Consequently, these types of mixing issues can be bypassed by acknowledging a *hierarchy* of discourse domains, some of which are fundamental and pure and others are non-fundamental and potentially impure. For instance, assuming that there is a fundamental and pure subject matter of physics, we can define under it an impure sub-domain of aesthetical physics that deals, for example, with the aesthetical features of entities relevant to physics inquiry. The existence of such a mixed domain does not threaten the integrity of the fundamental domains of physics and aesthetics, for no overlapping of such domains is

[5] For example, whether the theoretical virtue of simplicity does not bear any aesthetical content remains unclear, and similar concerns emerge from instances of concepts like symmetry, coherence, elegance, and harmony, which can be argued to be both scientifically and aesthetically relevant.

forced by acknowledging that there can be non-fundamental domains encompassing content from multiple domains.

Aligning with the explanation, nothing prevents one from forming compounds of atomics, such as "*The Birth of Venus* is colorful and *The Birth of Venus* is beautiful," that address multiple subject matters, yet where the constitutive sentences themselves count as members of a single fundamental domain. In this sense, we are discussing subject matters as categories of thought and speech that display hierarchy and fundamentality relations. At the most fundamental level, we have subject matters that are likely primitive categories of thought and discourse on the basis of which we can form mixed subject matters of various sorts.

Aligning with the notion of there being fundamental subject matters of atomics, Lynch treats atomic propositions as *essentially* belonging to only one domain:

> What makes a proposition a member of a particular domain? The obvious answer: the subject matter it is about. [...] [W]e believe all sorts of different kinds of propositions: propositions about ethics, mathematics, about the sundries of everyday life. No one, presumably, will deny that these propositions concern not just different subjects, but fundamentally different subjects. [...] Propositions are the kind of propositions they are essentially; therefore, belonging to a particular domain is an essential fact about an atomic proposition. (2009, 79–80)

While this aligns with the theoretical desiderata of how fundamental domains bear unambiguous identities and determinate extensions, it ought to be clarified why alethic theorists prefer such desiderata in the first place.

Starting with the truth pluralists, such theorists are explicit in their commitment to discourse domains as an explanatory resource. For pluralist accounts of various sorts, different kinds of sentences get to be true in distinct ways by possessing the operant truth-determining property for their domain:

> According to the alethic pluralist, there will be a robust property in virtue of which the propositions expressed by sentences in a particular domain of discourse will be true, but this property

> will change depending on the domain we are considering" (Edwards 2011, 31).[6]

According to pluralists, there is one-to-one correspondence between the domains of sentences, like physics and aesthetics, and distinct truth-determining properties, like correspondence and coherence. Subsequently, knowing the domain membership of all truth-apt sentences enables the pluralists to account for their truth or falsity in a domain-reliant manner by looking at whether any particular sentence possesses the truth-determining property for their domain.

Now, combine this idea with ambiguous or indeterminate accounts of domains, where there would be a range of truth-apt sentences, the domain membership of which would be confused. If there is a range of truth-apt sentences (S_1,…, S_n) that have confused membership between the domains of D_1 (physics/realist discourse) and D_2 (aesthetics/anti-realist discourse) with distinct truth-determining properties of P_1 (correspondence) and P_2 (coherence), the pluralists would be unable to account for their truth in a domain-reliant manner. For example, if it is not clear whether the sentence "*The Birth of Venus* is symmetrical" belongs to the domain of physics, which deals with the extensional composition of perceivable objects (physical symmetry), or to the domain of aesthetics, which deals with the projected phenomena of beauty and harmony (aesthetical symmetry), and where the respective domains are governed by distinct truth-determining properties of correspondence and coherence, then the pluralists would be unable to account for the truth or falsity of such sentences in a domain-reliant manner. Simply put, a question emerges as to what property such sentences must possess to count as true. No simple answer is forthcoming.

Perhaps such sentences count as members of both domains? The problem with this explanation is that according to the pluralists, what constitutes truth for each sentence is the possession of the distinct truth-determining property for its

[6] Without domains, explaining why any particular sentence is true on the basis of some specific factor, like coherence or correspondence, and not others becomes difficult if not impossible (Wyatt 2013, 231–232). Even worse, without domains, some sentences can be both true and false in some pluralist frameworks, conflicting with the law of bivalence and arguably even the laws of non-contradiction and identity (Edwards 2018b, 85–86; Tauriainen 2021, 198–199).

domain, and where its falsity lies in the lack of said property. If there is a range of sentences that count as members of multiple domains with distinct truth-determining properties, then such sentences can have one of the relevant truth-determining properties and lack another. Therefore, such sentences would have truth-determining and falsity-determining properties, rendering them simultaneously true and false, and thus conflicting with the standard law of bivalence, where each truth-apt sentence is either true or false but never both.

Another option would be to argue that such sentences belong to no domain, which would prevent the pluralists from accounting for the truth of such sentences in a domain-reliant manner. This would conflict with the basic pluralist permissiveness regarding truth-aptness: "Truth pluralists take the demands for truth-aptness to be very minimal, and focus their attention on what kind of truth a sentence is apt for" (Edwards 2018b, 95). Such permissiveness is a central motivation in arguing for the benefits of pluralist accounts over the traditional monist views that face pressing issues of scalability: "The most common pluralist move against monism is to invoke the so-called scope problem: no monist theory has a scope sufficiently wide to accommodate all truth-involving discourse" (Ferrari, Moruzzi, & Pedersen 2020, 631). Even worse, dispensing with the truth-aptness of the aforementioned types of sentences would generate truth value gaps, and thus necessitate inference with such a sentence's problematic, assuming the standard Tarskian definition of validity. Finally, the inability to account for the domain membership of confused sentences would render the pluralists' accounts incomplete, especially when we ordinarily take sentences like "*The Birth of Venus* is symmetrical" as capable of being true or false.

While critics like David (2013, 49; 2022, 8.2) have made skeptical remarks about the possibility of forming a sufficiently well-individuated account of discourse domains for the truth pluralists to scale their accounts, a neglected fact is that the substantive monist accounts also rely on there being

precise boundaries between discursive contents to demarcate truth-apt and non-truth-apt contents.[7]

Under substantive monist accounts, there is only one way of being true overall via sentences possessing the relevant truth property, such as correspondence or coherence. However, such accounts should explain why some sentences are susceptible to their preferred criteria for being true, which involves separating them into truth-apt and non-truth-apt domains. Separating sentences into such domains has been an important part of the historically significant debate between the classical or neo-classical correspondence and coherence theories, where the former accounts have difficulties in explaining the truth of ethical and aesthetical sentences, whilst the latter have difficulties in explaining the truth of physics and other natural sciences that are rendered so by mind- and theory-independent facets of the world.

More specifically, if there is no clarity on whether a range of sentences (S_1, …, S_n) belongs to D_1 (truth-apt) or D_2 (non-truth-apt) or both, then the monists cannot account for the truth-aptness of such sentences. Utilizing the same example sentence, "*The Birth of Venus* is symmetrical," which arguably belongs to both domains of physics and aesthetics or discourses about mind-independent and -dependent aspects of the world, at least some correspondence and coherence theorists would face difficulties in accounting for the truth-aptness of such sentences. Of course, monist theorists can deploy similar strategies to that above in dealing with such sentences, treating them as simply confused or removed from the range of truth-aptness, but they are also susceptible to similar definitional issues that follow.[8]

Simply put, with insufficient accounts of domains and their membership at hand, pluralists face challenges in accounting for the domain membership and, subsequently, the

---

[7] For example, David notes, "I want to remark in passing that the notion of a domain of discourse may well be a serious liability for pluralism about truth [...] Pluralism wants to sort propositions into different domains according to the subject matter they are about [...] Giving a principled account of how this is to be done is likely to be difficult" (2013, 49).

[8] Monists can reject the truth-aptness of such sentences more easily than the pluralists, for the former are not in general pushed towards accepting permissiveness regarding truth-aptness.

truth-value of *some* sentences, and monists face similar difficulties in demarcating the truth-aptness and non-truth-aptness of *some* contents. Likely, similar problems emerge for other philosophical theories as well, but their articulation is left to another occasion. We thus move to discuss the prospect of defining domains in a manner that posits them as having unambiguous identities and determinate rules for membership.

4. Individuating domains and accounting for their membership

The view of discourse domains that has been articulated thus far treats domains as classes of atomic sentences individuated by topical subject matters, where we further recognize a hierarchy of fundamental and non-fundamental subject matters and domains. For instance, a fundamental subject matter like physics or aesthetics would be a primitive category of discursive contents where a sentence counts as an instance of such a subject matter by composing of concepts of the aligning kind. Further, concepts fall under the aligning kinds on the basis of them denoting distinctively physical or aesthetical phenomena, or perhaps on the basis of them advancing discourse about relevant subject matters. In this sense, the sentence "snow is white" would be a distinctively physical sentence owing to its singular term concept referring to a range of extensional objects and where the predicate concept denotes an objective color property. Further, a non-fundamental domain would be such that it combines contents from two or more fundamental subject matters like physics and aesthetics.

As noted in the previous section, initially, one might wonder why one should bother to individuate fundamental subject matters and domains on topical rather than *ontological* grounds. There are several reasons for this. First, topical categories are widely utilized both within and outside of philosophy. Ordinarily, we take subject matters and discourse domains as primary topical categories, like physics and aesthetics, and this also aligns with our formal understanding of the world, where scientific disciplines are sorted into aligning domains of inquiry. Second, it is customary to hold that sentences addressing distinct subject matters, like physics and

aesthetics, can instantiate concepts denoting entities with varying ontological statuses. Nothing prevents the subject matter of physics from encompassing sentences that make reference to abstract objects or projected properties. Similarly, aesthetical sentences can concern objective properties of extensional objects, like whether some artwork qualifies as a mosaic. In this sense, there are well-grounded reasons for thinking that our mundane and theoretical ways of demarcating discursive content kinds are independent of concerns about the ontological status of entities denoted by the concepts that compose such a discourse. Third, it seems difficult to achieve a clear distinction between topical and ontological categories in the first place since there are abundant mutually exclusive ontological categories, and arguments for which of these are philosophically tenable or should be treated as fundamental are notoriously difficult to solve. Finally, as ontological categories are ultimately human categories and relative to the background theory through which they are formulated, this allows them to be treated as proto-subject matters or proto-topics, which results in the further blurring of boundaries between topical subject matters and ontological categories. Therefore, it is not even remotely clear whether ontological categories would provide any more robust distinction than topical categories for demarcating kinds of contents, especially when both categorizations are dependent on mind-dependent factors.

While both intuitively appealing and theoretically justifiable, the topic-based understanding of subject matters involves the cumbersome task of categorizing their contents like truth-apt atomics into the aligning domains. As noted, we can assume that this categorization happens at the level of singular term and predicate concepts of atomics. From this, we approach the question of demarcating singular term and predicate concepts into fundamental domains on topical grounds.

One problem is that there is no shortage of natural language concepts that can be deployed in the singular term or predicate positions of atomics, and assigning each of them to some topically individuated domain poses a challenging task that is subject to skeptical remarks. Utilizing the aforementioned example, it is not clear whether symmetry would be a concept or property that belongs to the domain of inquiry

about physics, mathematics, or aesthetics, or whether the predicate "is a mosaic" is a distinctively aesthetical concept when it concerns the material composition of extensional objects. Were one to argue that the concept of symmetry is ambiguous owing to the different ways of being symmetrical, an argument is required to explain why the phenomenon of symmetry is such that is permits a clear-cut disambiguation where nothing more than either physical, mathematical, or aesthetical symmetry is involved in each deployment of this concept. In relation to this, one might justifiably argue that in certain instances, the concept of symmetry denotes a property or phenomenon that is *simultaneously* relevant for both physics and aesthetics, and where these senses cannot be straightforwardly separated. We will discuss the problems caused by such mixed concepts further in the following section.

Another challenge in achieving a well-individuated account of topical domains, and of ontologically individuated domains for that matter, follows from the fact that our conceptual frameworks change, as do our conceptions of what the identities of subject matters are, which of them are fundamental, and what concepts instantiate which subject matters. This is also true for ontological categories that are subject to change according to the development of our metaphysical understanding of the world. For instance, while one could argue that there is a distinctively psychological domain that deals with discourse about mental states and experiences, nothing in principle prevents our conceptual frameworks from changing in a way that reduces this domain to one concerning a simple material change of complex systems—that is, physical and chemical processes of the brain. In such a hypothetical instance, what is now considered its own distinctive domain of psychology with its distinctive concepts would eventually reduce to a more fundamental domain, thus posing a challenge for providing robust accounts of domains that would persist over time by rendering one's account of their individuating factors relative to the present time and the contingent conceptual framework from which the individuating distinctions are drawn and justified. Further relativization would follow from there being competing theories or frameworks of thought that can provide incom-

patible understandings of what subject matters exist, what their boundaries are, and which of them are fundamental.

Hence, the topics-based approach, while intuitively appealing, suffers from a general lack of facts for grounding precise boundaries between subject matters and their respective concepts, thereby casting suspicion on the ability to form a well-individuated account of discourse domains when individuated on topical grounds.

However, this feature of our conceptual frameworks being subject to endless progress and re-evaluation concerns almost *all* philosophical theories, and as such provides a poor critique of one's account of domains per se. Similarly, such a conclusion does not diminish the prospect of there being better or worse ways of defining domains relative to each theoretical context or conceptual framework, and it thus allows the possibility that domains can be defined as well-individuated classes relative to the assumed background theory or conceptual framework.

It would also be apt to further discuss the actual ways in which discourse domains can be defined relative to the assumed background framework. In addition to the described topic-based and ontology-based approaches to individuating subject matters, by articulating a promising view that accounts for the domain membership of sentences via the *functional* or *teleological* role of their constitutive concepts, we can discuss problems with such accounts in the following section.

5. Functional or teleological approach to identifying subject matters and discourse domains

According to the functional or teleological view, truth-apt atomics are categorized into kinds according to the functional roles of the relevant constitutive concepts of truth-apt contents:

> The suggestion is that we can individuate kinds of predicates in accordance with the general functional roles that those predicates are taken to have. These are intended to mark fairly intuitive distinctions between kinds of subject-matter" (Edwards 2018a, 63; cf. Gemester 2020, 11353).

Note that Edwards, for one, accounts for the domain membership of atomics by kinds of concepts, where the consideration is restricted to the predicate concept: "So, it is not what a sentence is about that we should be considering for domain membership, it is rather how the thing the sentence is about is represented, by the use of a predicate to attribute a property" (2018b, 96; cf. Pedersen & Wright 2018, 4.5). The reasoning is relatively straightforward: atomic sentences are always *about* the objects designated by or referred to by singular terms, but what renders such sentences bearers of content is that something *is said* about these objects in the form of predication.[9] A less controversial claim would be that predication is what renders atomic sentences truth-apt, and hence the predicate concept should be taken as the *primary* content kind when considering the domains of atomics.

Ferrari promotes a view along these lines, arguing that a singular term can sometimes help disambiguated ambiguous predicates and hence have a secondary role in assigning atomics into domains: "However, looking at the predicative expression may not always be enough to determine to which domain a proposition belongs. When this is the case, we need to look also at the main subject matter of our judgement" (2021, 33). For instance, in the case of ambiguous predicates that potentially assign sentences to the distinct domains of personal taste and ethics, like "is good," the respective singular terms of "sushi is good" and "charity is good" help to disambiguate the initially ambiguous predicates and assign the sentences to the appropriate domains. Evidently, this is in stark contrast to Edwards, according to whom "Atomic sen-

[9] Edwards motivates the predicate-emphasizing approach to domain membership as follows: "I will suggest that it is the predicate that determines the domain [of atomic sentences]. We can distinguish between two things: what a sentence is about, and what is said about the thing the sentence is about. A sentence is about its object […] But what makes these things sentences is that there is more: there is something that is said about the things that the sentences are about. […] It is this aspect—the attribution of a property to an object—that makes these kinds of sentences *sentences* in that they are bearers of content. So, it is not what a sentence is about that we should be considering [when assigning them into domains,] it is rather what is said about the thing the sentence is about" (2018a, 78–79).

tences are thus assigned to domains by the predicate they contain. The singular term is not relevant to domain individuation" (2018b, 97). However, even if it is controversial, we can simply accept the predicate-emphasizing approach since demonstrating problems with this strategy also scales to more complex strategies that look at *both* the singular term and the predicate concepts when accounting for the domain membership of atomics.

Assuming this premise and returning to the case of functional analysis, predicates like "is a proton" are distinguished as distinctively physical owing to their ability to "mark features of fundamental phenomena, such as matter, mass, and force," and predicates like "is beautiful" are distinguished as aesthetical owing to their ability to "mark a particular kind of the sensory features of an object" (Edwards 2018a, 66). While this is not the place to provide an extensive analysis of the philosophical sustainability or strengths and weaknesses of such an approach, there are a few skeptical notes that can be made to demonstrate that even this strategy does not offer a confusion-free method of individuating discourse domains.

First, the functionalist strategy relies on existing taxonomical distinctions (i.e., subject matters) between discursive contents to allow for categorizing their functional roles into the kinds articulated above. To be able to define the functional role of "is white" as an aesthetical rather than a physical predicate, some pre-existing distinctions for distinguishing between such predicate kinds ought to be in place. Further, defining such pre-existing proto-distinctions or subject matters would lead to similar issues with defining topical (or ontological) subject matters. Therefore, and partially due to the need for there to be prior taxonomical distinctions to define the functional roles of predicates, the functional strategy is susceptible to fringe cases where the domain membership of atomics would be unclear due to the presence of instantiating predicates that encompass confused content or bear mixed functional roles.

Second, Edwards (2018a, 81) acknowledges that what determines the domain membership of atomics is the *primary* functional role of predicates. However, this implies that predicates can also have secondary functional roles, which creates the need to offer some account for distinguishing such roles

in the case of any particular predicate. Again, in the case of "is white," such a predicate can have the primary functional role of advancing aesthetical discourse on one occasion and a physical role on another occasion without any clear prospect for distinguishing between such roles beforehand. Hence, the predicate would be able to assign one and the same sentence to distinct domains depending on the instance that determines its primary functional role, which would potentially result in issues where sentences have either confused domain membership or belong to multiple domains between instances. This is merely intuitive, for nothing prevents a single predicate from advancing discourse about *both* physics and aesthetics, yet it is difficult to see how a scaling account can be offered for determining which type of discourse is primarily being advanced in any particular instance, especially when keeping in mind the already discussed feature of our conceptual frameworks being susceptible to constant development and change. Therefore, while my contention is that this does not render the functional approach inherently flawed or necessarily more problematic than the alternative views, this approach does not provide an unproblematic foundation for defining domains as unambiguous classes of sentences with determinate rules for membership.

6. Complex content

In addition to the aforementioned problems in defining subject matters and achieving a well-individuated account of discourse domains on topical, ontological, and functional or teleological grounds, there are neglected issues with complex content that compromise one's ability to define domains as unambiguous classes of sentences with determinate rules for membership under all of the aforementioned strategies.

Starting with the problem of ambiguity, insofar as discourse domains are defined for a natural language *L*, then the inherent ambiguities involved with such languages risk being transferred to one's account of domains. Natural languages encompass polysemous terms that can allow for multiple and mutually incompatible readings, and this lays the foundation for the phenomenon of lexical ambiguity to emerge, where the meanings or referents of terms can be confused. From the

phenomenon of lexical ambiguity emerges semantic ambiguity, where sentences composed of ambiguous terms allow for multiple and mutually exclusive readings in a potentially indeterminate or confused manner. The problem that such ambiguity poses for one's account of domains is that insofar as sentences are assigned to domains on the basis of the concepts deployed in the singular term or predicate positions, yet where both the singular term and predicate terms can encompass ambiguity, then our ability to assign such sentences to domains in an unambiguous manner will be compromised even if the respective (fundamental) domains themselves have unambiguous identities. For example, ambiguous predicates, such as "is white," compromise one's ability to assign sentences to a single domain in a determinate manner according to the predicate allowing for *both* objective color-property and projected social-property readings, which would assign the respective sentence to the independent discourse domains of physics, sociology, and perhaps even aesthetics. Note that the initial solution proposed by Ferrari (2021), where a singular term can help to disambiguate an ambiguous predicate, does not work in full scale since predicates, like "is white," can apply to the same unambiguous or ambiguous singular term. Similarly, while a functional analysis can help to disambiguate such predicates in *some* contexts, nothing prevents instances where confusion persists between, for example, the primary and secondary functional roles of such predicates.

However, the aforementioned ambiguity issues are well known and there are effective methods for philosophers to deal with them from both theoretical and pragmatic perspectives (Sennet 2021). Theoretically, perhaps the most efficient way of dissolving lexical and semantic ambiguities is to *not* treat sentences as the contents of domains. Rather, one can adopt *interpretations* of atomic sentences or atomic proposition as the contents of domains to avoid issues of lexical and semantic ambiguities. While sentences like "Charlie is white" are ambiguous because they allow for multiple interpretations in an indeterminate manner, the interpretations themselves have, at least when casting vague expressions outside the range of consideration, clear and determinate meanings. In this sense, ambiguous sentences allow for multiple readings, yet these readings themselves are what cognitive agents

are able to clearly and unambiguously identify. Simply put, in the case of the aforementioned predicate, one always understands "is white" as *either* a physical, aesthetical, or social predicate, and there is arguably no confusion between these distinct interpretations as they display clear variance in the kind of their content. Assuming that one has a well-individuated account of domains and robust rules for membership, then any interpretation of an atomic sentence would determinately assign the sentence to an appropriate domain independently of our ability to identify definitively whether any particular sentence or concept stands for this or that reading. From this, it follows that the issue of ambiguity can be constrained wholly to the side of the language or our ability to *know* which sentences should be interpreted in what ways. Hence, this can occur for the language for which one defines domains, and it does not threaten the prospect of reaching a well-individuated account of domains and their membership for disambiguated contents, such as interpretations of sentences or concepts.

Beyond the theoretical prospects of satisfactory disambiguation, there are also effective ways of dissolving language-bound ambiguity on pragmatic grounds, and thus of reaching a more desirable account of discourse domains overall. In general, the problems caused by ambiguity can be managed by *regimenting* the discursive contents over which domains scale. For example, in certain technical contexts where ambiguity regarding our ability to know about the domain membership of atomics can cause issues, one can simply eliminate ambiguous terms or disambiguate them by adding indications for correct readings. In this sense, the predicate "is white" can be disambiguated to encompass two distinct readings, "is white [in color]" and "is white [in class]," encompassing distinct content kinds and having their own application rules, subsequently governing membership to the respective domains of inquiry of physics and sociology.

However, it is worth emphasizing that regimenting the whole range of natural discourse will not do. One reason for this is that polysemy-based ambiguity is a feature, not a bug, of such discourse. In general, polysemous and ambiguous discourse can be useful, where we sometimes *want* our speech to be confused. For instance, when we watch improvi-

sation theater or read a piece of literature, we do not mind that expressions sometimes allow for multiple readings in an indeterminate manner. There is also strategic ambiguity, for example, in the case of the US–Taiwan situation, where the United States' commitment to defend Taiwan from possible invasion from foreign forces is left intentionally ambiguous for political purposes. Nonetheless, unregimented natural discourse can be allowed to encompass these types of ambiguities, under which our knowledge of the domain membership of some sentences is subsequently confused, yet membership-governing concepts can be appropriately regimented in technical contexts of various sorts that benefit from there being precise and known boundaries between discursive contents.

Based on the aforementioned discussion, ambiguity does not provide a serious threat to reaching a philosophically sustainable account of discourse domains. However, beyond ambiguity, there is a distinct and neglected phenomenon of complex and *mixed* content that poses a threat to reaching a well-individuated account of discourse domains even after following the disambiguating strategies for both the topical and functional strategies. For the sake of argument, we can assume that an unambiguous account of discourse domains can be achieved by restricting the contents of such domains as interpretations of atomics, where each interpreted predicate concept assigns sentences to only one topically individuated fundamental domain. From this, we reach the question of whether all disambiguated domain-relevant predicate concepts are such that they are governed by or fall under only one fundamental domain, or whether all such concepts address only one primary subject matter. Aligning with the intuition that some concepts address multiple subject matters at once, and hence govern membership to more than one domain, nothing in principle prevents there being mixed concepts that encompass content that is equally relevant to multiple fundamental subject matters at once or that advance discourse about distinct subject matters on equal grounds.

An example of mixed content would be a concept or sentence encompassing content from multiple topically individuated domains at once or when abiding by the functional approach that simultaneously advances discourse about more

than one subject matter with no prospect for separating primary and secondary functional roles. It is worth emphasizing that here, the focus is strictly on predicate concepts and their mixing, but matters are only complexified if singular terms are allowed to govern domain membership, since nothing prevents them from being mixed as well.

Returning to the previously introduced example case of symmetry, the problem with this concept is that it arguably not *only* presents ambiguity between physical, mathematical, and aesthetical readings or ways of being symmetrical but, as a phenomenon or property of both concrete and abstract objects, is complex enough to warrant a view where it can compose of content that is relevant for multiple subject matters simultaneously. This is because, in certain instances, it is reasonable to hold that symmetry denotes a phenomenon that encompasses physically, mathematically, and aesthetically relevant content. For example, nothing prevents thinking that in the case of certain natural symmetries, like the fractal structures of snowflakes, the phenomenon of symmetry is inseparably physical and mathematical, or when discussing the symmetry of an artwork, like architectural elements, there can be inseparable physical, mathematical, and aesthetical content involved. In this sense, there are reasons to believe that in some cases, symmetry as a concept denotes a phenomenon encompassing content that is relevant to more than one fundamental subject matter or domain of inquiry based on, for example, it concerning the harmony or balance of portions of concrete and abstract objects of various forms in a sense that is relevant to physical, mathematical, and aesthetical domains of inquiry. Such a balance of portions can be an extensionally manifesting natural phenomenon of material objects and can sometimes even act as a precondition for certain biological processes to emerge. Further, such a balance of portions is an inseparable component of the phenomenon of aesthetical symmetry, which concerns the perceived symmetricity of concrete or abstract entities. However, while both physical and aesthetical symmetries are such that no criterion of idealization is required, in at least some of the mathematical senses of symmetry, only theoretical entities displaying a perfect or idealized balance of portions count as symmetrical, where some such symmetries cannot even, in

principle, manifest in extensional objects, and where other mathematically symmetrical objects might be infinite to the point of inconceivability, hence repelling evaluations of aesthetical symmetry.[10] In this sense, at least certain theoretical conceptions of symmetry require from symmetrical objects more than a simple balance of portions by, for example, requiring a symmetrical object to display an idealized property of being perfectly symmetrical. This complexifies matters by raising a concern about the concept of symmetry being able to denote distinct *kinds* of symmetries, rendering the general-level concept potentially ambiguous between such kinds but also rendering some of these kinds mixed, where they inseparably involve content that is relevant for more than one fundamental subject matter at once. Consequently, truth-apt sentences bearing the concept of symmetry pose a troublesome case for assigning them to fundamental domains in an unambiguous and determinate manner.

Here, the skeptic might contend that such instances are not really about the concept or property of symmetry being mixed between topical subject matters but rather demonstrate that the general-level concept of symmetry is simply ambiguous regarding the different ways of being symmetrical. To emphasize, however, the point here is not that symmetry as a concept is *only* ambiguous between different readings but that the concept of symmetry can sometimes denote phenomena that are relevant for multiple subject matters at once. In this sense, the question of whether some object is physically symmetrical can, in some instances, be *inseparable* from whether it is also aesthetically symmetrical, or the phenomenon of physical symmetry can be inseparably entwined with a mathematical understanding of symmetry. Therefore, and aligning with the intuition of how some concepts can bear content of distinct kinds or address or fall under multiple fundamental subject matters at once, it is a reasonable assumption that even after conducting the disambiguating programs, all domain-membership governing concepts do not assign truth-relevant contents to only one fundamental domain.

[10] My contention is that inconceivability does not preclude aesthetical evaluations.

Moreover, some such concept can assign contents to distinct domains that can either be truth-apt and non-truth-apt or susceptible to being true in different ways, raising concerns about the subsequent definitional issues for some monist and pluralist theories of truth, which were discussed previously.[11] Simply put, insofar as a monist would argue for the truth-aptness of physical discourse while rejecting the truth-aptness of aesthetical discourse, sentences like "School of Athens is symmetrical" can prove problematic by falling within both the aforementioned domains. Similarly, the truth pluralists face difficulties in articulating the way in which such sentences get to be true when the aforementioned domains are governed by distinct truth- and falsity-determining properties, and where the respective sentence can possess one of these properties while lacking another.

7. Concluding remarks

From a definitional standpoint, both alethic monists and pluralists would benefit from fundamental domains over which truth-aptness or ways of being true vary as unambiguous classes of sentences with determinate rules for membership. However, a few considerations in this paper have aligned with one another to formulate a joint argument against the idea of fundamental discourse domains as such classes when individuated on topically understood subject matters.[12] First, the project of defining subject matters as well-individuated categories is susceptible to indeterminacy due to the general lack of boundaries for demarcating content kinds on particular grounds and where such boundaries are unstable over time owing to inevitable development and changes in our conceptual frameworks through which the deployed topical distinctions are justified. Second, even after deploying certain

[11] The question of whether mixed concepts are vague concepts or what their relation is to one another ought to be addressed in full detail in an independent study. My contention is that mixed content is distinct from vague content since the former can enable a clear compositional analysis.

[12] All this leaves open whether the desired account of domains could be achieved on ontological rather than topical grounds. However, because of the extensiveness of this topic, ontology-based approaches to discourse domains ought to be examined elsewhere in detail.

disambiguating strategies, defining fundamental domains as well-individuated classes of sentences faces problems due to the intuition that not all concepts are simple or univocally about a single subject matter. According to the example case of mixed content, some concepts can encompass content that is simultaneously relevant to multiple subject matters, which are relevant to more than one domain of inquiry or advance multiple discourse at once without any prospect of precisely separating primary and secondary functional roles. Insofar as such concepts compose atomics, which are responsible for domain membership, such sentences can arguably belong to more than one domain at once. This causes problems for alethic theorists who bind truth-aptness or distinct ways of being true to domains rather than individual sentences.

Therefore, the concluding argument is that alethic theorists and others who rely on natural language discourse domains as an explanatory resource should consider a commitment to moderate indeterminism about the extensions of fundamental discourse domains when individuated on topical grounds, where general guidelines for assigning sentences to domains can be provided, yet where the domain membership of *some* sentences cannot be unambiguously accounted for. At the bare minimum, such a conclusion is in stark contrast to Lynch's (2009, 79–80) early approach and Edwards' (2018a, 79; 2018b, 96–97) current approach that insist on atomics belonging solely to one fundamental domain. The promoted view also contrasts with those for whom truth-apt atomics can belong to multiple domains yet where determinate rules for primary domain membership can nonetheless be given (Wyatt 2013, 233). In conclusion, insofar as one argues that all domain-relevant concepts ought to be defined in a manner that posits them as being governed by only one (primary) fundamental domain, then such theorists should address the neglected issues that complex and mixed contents pose for assigning all truth-apt content to discourse domains in an unambiguous and determinate manner.

University of Jyväskylä

References

David, M. (2013). "Lynch's Functionalist Theory of Truth." In N. J. L. L. Pedersen and C. D. Wright (eds.), *Truth and Pluralism: Current Debates.* Oxford: Oxford University Press, 42–68.

David, M. (2022). "The Correspondence Theory of Truth." In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 edition).

Edwards, D. (2011). "Simplifying Alethic Pluralism." *The Southern Journal of Philosophy* 49(1), 28–48.

Edwards, D. (2018a.) *Metaphysics of Truth.* Oxford: Oxford University Press.

Edwards, D. (2018b.) "The Metaphysics of Domains." In J. Wyatt, N. Pedersen, and N. Kellen (eds.), *Pluralisms about Truth and Logic.* London: Palgrave Macmillan, 85–106.

Ferrari, F. (2021). *Truth and Norms: Normative Alethic Pluralism and Evaluative Disagreements*. London: Lexington Books.

Ferrari, F., S. Moruzzi, and N. Pedersen (2020). "Austere Truth Pluralism." In M. Lynch, J. Wyatt, J. Kim and N. Kellen (eds.), *The Nature of Truth* (Second edition). Cambridge, MA: MIT Press, 629–656.

Gemester, W. (2020). "Shopping for Truth Pluralism." *Synthese* 198(12), 11351–11377.

Kim, S. and N. Pedersen (2018). "Strong Truth Pluralism." In J. Wyatt, N. Pedersen, and N. Kellen (eds.), *Pluralisms in Truth and Logic.* London: Palgrave Macmillan, 107–131.

Lynch, M. (2009). *Truth as One and Many*. Oxford: Oxford University Press.

Pedersen, N. and C. Wright (2018). "Pluralist Theories of Truth." In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 edition).

Pedersen, N., J. Wyatt, and N. Kellen (2018). "Introduction." In J. Wyatt, N. Pedersen, & N. Kellen (eds), *Pluralism in Truth and Logic.* London: Palgrave Macmillan, 3–35.

Sennet, A. (2021). "Ambiguity." In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 edition).

Tauriainen, T. (2021). "No Safe Haven for Truth Pluralists." *Acta Philosophica Fennica* 97, 183–205.

Wyatt, J. (2013). "Domains, Plural Truth, and Mixed Atomic Propositions." *Philosophical Studies* 166(1), 225–236.

Is Logic Normative?

ANANDI HATTIANGADI

1. Introduction

Though it is hardly uncontroversial, the thesis that logic is normative enjoys widespread agreement—probably just about as much agreement as one is ever likely to find in philosophy.[1] There is far less agreement, however, on what exactly this thesis amounts to. To begin with, proponents of the thesis can't seem to agree on whether the normative authority of logic is *robust* or *weak*.[2] If logic is robustly normative, it has a normative authority that is independent of our attitudes or conventions; if it is weakly normative, it has a normative authority that is entirely dependent on our attitudes or conventions. This fundamental disagreement about the normative authority of logic seems to leave little room for any point of agreement among the proponents of the thesis. Furthermore, some opponents of the thesis allow that logic is "entangled" with the normative to the extent that it has normative consequences that are instrumental to the achievement of our wider goals (Russell 2017). This makes it difficult to discern

[1] Proponents of the thesis include Ayer 1946; Ayer et al. 1936; Carnap [1937] 2001; Beall & Restall 2006; Caret 2016; Frege [1897] 1997; Field 2009a; 2009b; 2009c; 2015; Kant [1800] 1974; Keefe 2014; Pettigrew 2017; Priest, 1979; Railton 2000; Read 2006; Sainsbury 2002; Steinberger 2017b, 2019; Warren 2020; Woods 2023. Opponents include Harman 1986, Russell 2017, and Pigden and Olsen, ms.

[2] Though the issue is not always taken up explicitly, those who seem to hold that logic is robustly normative include Frege [1897]1997 and Kant [1800] 1974. Those who hold that logic is merely weakly normative include Ayer 1946; Carnap [1937] 2001; Field 2009a, 2009b, 2009c, 2015; Warren 2020; and Woods 2023.

any daylight between the views of those who hold that logic is not normative and those who hold that it is only weakly so.

In the next section, I will argue that the thesis that unites the proponents and excludes the opponents is that logical statements and the judgments they can be used to express—such as those concerning logical validity or logical entailment—*are* normative statements and judgments, in the sense that they analytically, semantically, or conceptually have normative consequences. In section 3, I will critically assess whether logical statements and judgments are indeed normative in this sense. I will consider the prospects of various accounts of what the normative consequences of logical statements or judgments might be, and find them all to be wanting. This, I claim, gives us good reason to deny that logic is normative.

2. What is at issue?

To discover what is fundamentally at issue in debates about the normativity of logic, it will be helpful to consider the fault lines and alliances among the various parties to the debate.

First, there is the "absolutist" view, handed down from Frege and Kant, according to which logic is robustly normative. Kant, for instance, characterized logic as consisting of "the absolutely necessary rules of thought" (A52/B76), which instruct us not "how the understanding is and thinks" but "how it *ought* to proceed" (Kant 1800/1974, 16; quoted in Steinberger 2017a). Frege, in a similar vein, says the following:

> Just as 'beautiful' points the way for aesthetics and 'good' for ethics, so do words like 'true' for logic…When we speak of moral or civil laws, we mean [meinen] prescriptions, which ought to be obeyed but with which actual occurrences are not always in conformity. Laws of nature are general features of what happens in nature, and occurrences in nature are always in accordance with them. It is rather in this sense that I speak of laws of truth. Here of course it is not a matter of what happens but of what is. From the laws of truth there follow prescriptions about asserting, thinking, judging, inferring. (Frege 1918/1997, 325)

As I read this passage, Frege favourably compares logic to the paradigmatically normative disciplines of ethics and aesthetics. He goes on to consider whether logical laws—the laws of truth—resemble more closely the laws of physics or the laws of morality. His answer is that they are a bit like both.[3] On the one hand, the laws of truth resemble the laws of physics in being objective, albeit "not a matter of what happens, but of what is." On the other hand, the laws of truth resemble moral laws in giving rise to "prescriptions about asserting, thinking, judging, inferring." Elsewhere, Frege describes logic as "a normative science", the aim of which is to prescribe "rules for our thinking and for our holding something to be true" (Frege, 1897/1997, 228). In a nutshell, absolutists hold that there is one true logic that reflects the normative *facts* regarding how we ought to think or reason.

In the early part of the 20th century, logical conventionalists repudiated the absolutist conception of logic as unscientific (Ayer 1946; Ayer et al. 1936; Carnap [1937] 2001). Yet, they nonetheless held on to the view that logic is normative. They sought to naturalize the normativity of logic by casting it as a product of our practices, as more like the laws of the state than the laws of nature. Ayer puts the point as follows:

> …what are called a priori propositions do not describe how words are actually used but merely prescribe how words are to be used. They make no statement whose truth can be accepted or denied. They merely lay down a rule which can be followed or disobeyed. Their necessity then, we must say, consists in the fact that it does not makes sense to deny them. If we reject them we are merely adopting another usage from that which they prescribe. (Ayer et al. 1936, p. 20)

---

[3] Glüer and Wikforss (2009, 65) take this passage from Frege to show that he held that logic is not normative. As they see it, Frege distinguishes the laws of logic from *both* the laws of nature and the laws of the state, treating the laws of truth as *sui generis*. However, this reading of Frege does not explain the final sentence quoted above, in which he says "from the laws of truth there follow prescriptions," nor does it explain why he says: "Just as 'beautiful' points the way for aesthetics and 'good' for ethics, so do words like 'true' for logic" (Frege 1918/1997, 325). I am grateful to Alex Miller for discussion on this point.

Moreover, Ayer goes on to say that the choice of a logic is in a sense arbitrary, since we could have chosen to adopt different conventions (Ayer et al. 1936, 21). Carnap echoes both Ayer's claim that the logical laws are in a sense up to us, and that this allows for a plurality of logical systems, since there are no normative, logical facts to be discovered:

> *In logic there are no morals*. Everyone is at liberty to build his own logic, i.e. his own language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments. (Carnap [1937] 2001, §17)

In saying that "there are no morals" in logic, I take Carnap to be denying the absolutist view that there are normative facts regarding how we ought to or are permitted to reason, and in saying that we should "give syntactical rules," he is implicitly committing to the weak normativity of logic. After all, rules tell us what we ought to or are permitted to do. Contemporary conventionalists similarly view the adoption of a logic as fundamentally the adoption of a system of normative, logico-linguistic rules governing our use of logical terms (cf. Warren 2020; Woods 2023). Other scientifically minded philosophers have similarly endorsed the view that logic is weakly normative (cf. Field 2009a, 2009b, 2009c, 2015).

At first blush, these two approaches to the normativity of logic seem to be too different to share a common core. Indeed, some debates about the normativity of logic concern the question whether logic is robustly normative, on which these two approaches disagree. However, there is a point of agreement between them: both are committed to the view that logical *statements* and *judgments* are normative. For instance, in the passages quoted above, Ayer says that "a priori propositions," including logical ones, "prescribe how words are to be used'," while Carnap encourages logicians to "give syntactical rules instead of philosophical arguments." Kant takes logic to consist in the "rules of thought," while Frege claims from logic "there follow prescriptions about asserting, thinking, judging, inferring." More recently, Field (2009a, 2009b, 2015) has argued that logic is normative in the sense that the concept of logical validity has a normative *role*, which cashes out as a policy regarding the formation and maintenance of

belief (Field 2015). Specifically, on this view, *what it is* to judge that an inference is valid is to have a policy of not believing the premises of the inference without believing its conclusion.

Critics of the thesis that logic is normative often point out you could read a whole textbook on logic without coming across paradigmatic normative terms such as "ought," or "may" anywhere (cf. Harman 1986). A logic is essentially a specification of a consequence relation on a set of truth bearers, so the core of a logic consists of statements of the following form, where $\vDash$ is a consequence relation, and P_1, …, P_n are the premises of an argument of which *C* is its conclusion:

(1) $P_1, \ldots, P_n \vDash C$.

Notably, there are no paradigmatic normative terms in statements of the form of (1). Statements of this form do not explicitly say anything about what one ought to do. Informally, logical statements include the following (with the key logical terms in italics):

(2) The Law of Excluded Middle is *valid*.

(3) "All ravens are black" *entails* "Ravi the raven is black."

(4) If the coin is either in the left hand or in the right hand, and the coin is not in the left hand, it *follows* that the coin is in the right hand.

Once again, none of these statements seem to be explicitly normative.

Now, it is highly unlikely that this point was lost on any of the proponents of the thesis that logic is normative. So, what could they have meant? Frege gives us a clue in the passage quoted above, when he says that prescriptions "follow" from the laws of truth, suggesting that logical statements are normative in virtue of having *normative consequences*. Similarly, it is possible to view proponents of the view that logic is weakly normative, such as Ayer, Carnap, and Field, as claiming that the concept of logical validity is a *thick normative concept*, much like "courage" or "greed," in that it has both a

descriptive content and a normative one.[4] On this view, an assertion of (3), for instance, would express a normative content, such as a policy not to believe "All ravens are black" while not believing "Ravi the raven is black." Thus, it seems plausible that what unites the proponents of the thesis that logic is normative is that logical statements and judgments have normative consequences.

However, this thesis does not exclude opponents of the normativity of logic. For instance, Russell (2017, 380), who holds that logic is purely descriptive, maintains that logic may be entangled with the normative to the extent that logical statements have normative consequences in conjunction with other normative statements, such as the statement that one ought to have only true beliefs, or that one ought to avoid reasoning to false conclusions. In light of this, Russell takes proponents of the thesis that logic is normative to be minimally committed to the claim that logical statements and judgments have normative consequences *all on their own* (Russell 2017, 379). But if logical statements and judgments have normative consequences all on their own, then these consequences must in some sense be analytic, semantic or conceptual. Indeed, many proponents of the normativity of logic explicitly commit to the analyticity of logical rules. According to conventionalists, the rules of a logic are thought to be analytic in the sense that they constitute the meanings of the logical constants and determine which inferences are valid (Ayer 1946; Carnap [1937] 2001; Warren 2020). Beall and Restall (2006) only count as admissible those precisifications

[4] Note that Field (2015, 55) claims that it would "sully the purity of logic to define validity in normative terms whose exact content is less than clear". Perhaps what he is expressing here is opposition to the *analysis* of the concept of validity in normative terms. After all, he very clearly accepts that validity has a normative *conceptual role*. For instance, he spells out the "conceptual role" of the concept of validity as follows: "To regard an inference or argument as valid is (in large part anyway) to accept a constraint on belief: One that prohibits fully believing its premises without fully believing its conclusion" (Field 2015, 42). Moreover, it is plausible that the concept of validity derives normative content from its normative role. If to regard an inference as valid *is* to accept a normative constraint on belief, then it is plausible that the concept of validity inherits the content of the constraint.

of "valid" that are formal, necessary, and normative. On this view, the normative consequences of judgments of validity are semantic or conceptual since they constrain which concepts count as concepts of validity. Thus, I understand the thesis that logic is normative to be the following:

> THE NORMATIVITY OF LOGIC (NL): Logical statements and judgments analytically entail normative consequences.

In the next section, I will test the plausibility of NL.

3. Are logical statements or judgments normative?

What might be the analytic normative consequences of our logical statements or judgments? This question may seem difficult to answer, given the large class of possible normative consequences our logical statements might have. However, it can be made more tractable by appeal to MacFarlane's (2004) taxonomy and nomenclature for normative bridge principles (cf. Steinberger 2016). These principles can be distinguished along several dimensions, as follows.

As I suggested above, I take the basic form of logical statements to be $P_1, \ldots, P_n \vDash C$. Now, let Φ be a normative operator (such as ought, may, or reason), $\mathbf{A}$ be an attitude operator (most commonly belief), and $\Phi(\mathbf{A}(P_1),\ldots,\mathbf{A}(P_n), \mathbf{A}(C))$ be a normative statement of some kind concerning changes of attitude (cf. Harman 1986; MacFarlane 2004; Steinberger 2016, 2017a). Then logical statements must have normative consequences of the following form:[5]

$$\Phi(\mathbf{A}(P_1), \ldots, \mathbf{A}(P_n), \mathbf{A}(C)).$$

For instance, if we let $\mathbf{A}$ stand for belief, and the normative operator to be "ought," we get the following bridge principle, where $\rightarrow$ stands for analytic, semantic or conceptual entailment:

---

[5] I am focusing here on the thought that logic is normative for theoretical reasoning, understood as reasoned changes in belief. This is by far the most common view of what logic is normative for among proponents of the normativity of logic (cf. Field 2009, Steinberger 2019). Some hold that logic is normative for our discursive practices (Dutilh-Novaes 2015), but I will set this view aside here.

$P_1, \ldots, P_n \vDash C$ → If *S* believes each of $P_1, \ldots, P_n$, then *S* ought to believe *C*.

Normative Operators: MacFarlane distinguishes between bridge principles which differ with respect to the deontic operators involved—ought (o), permission (p), or reason (r). For instance, (p) is a permissive principle, and (r) is a reason-involving principle:

(p) ($P_1, \ldots, P_n \vDash C$) → if *S* believes $P_1, \ldots, P_n$ then *S may* believe *C*.

(r) ($P_1, \ldots, P_n \vDash C$) → if *S* believes $P_1, \ldots, P_n$ then *S has a reason* believe *C*.

If the category of the normative is broadly construed, we might want to consider evaluative operators, such as "good" (g) and aretaic operators, such as "virtuous" (v) as well:

(g) ($P_1, \ldots, P_n \vDash C$) → if *S* believes $P_1, \ldots, P_n$ then it is good that *S* believes *C*.

(v) ($P_1, \ldots, P_n \vDash C$) → if *S* believes $P_1, \ldots, P_n$ then it is virtuous for *S* to believe *C*.

Scope: Bridge principles may differ with respect to the scope of the deontic operators, which can be narrow (C), Wide (W), or Distributed (D). For instance (Co) is a principle that involves the ought operator "o", and takes narrow scope, "C," while (Wp) takes wide scope and has the permissibility operator, and (Dr) employs the reason operator which is distributed over the conditional:

(Co) ($P_1, \ldots, P_n \vDash C$) → If *S* believes $P_1, \ldots, P_n$ then *S* ought to believe *C*.

(Wp) ($P_1, \ldots, P_n \vDash C$) → It may be the case that: if *S* believes $P_1, \ldots, P_n$ then *S* believes *C*.

(Dr) ($P_1, \ldots, P_n \vDash C$) → If *S* has a reason to believe $P_1, \ldots, P_n$ then *S* has a reason to believe *C*.

Polarity: Finally, bridge principles may differ with respect to the polarity of the belief in *C*.

Positive polarity (+). One ought to/may/has a reason to *believe C*.

> **Negative polarity (–).** One ought to/may/has a reason *not to disbelieve C.*

For instance, all of the above examples have had positive polarity. In contrast, (Wo–) takes wide scope over the ought operator and has a negative polarity.

> **(Wo–)** (P_1, …, $P_n \vDash C$) → It ought to be the case that: if *S* believes P_1, …, P_n then *S does not disbelieve C.*

I have argued previously that logical rules cannot be adopted (Hattiangadi 2023), following Kripke (forthcoming). This argument calls into question the very thought that such rules could play the kind of role in determining the meanings of logical terms that conventionalists suggest. Here, I set aside the question of whether it even makes sense to treat rules or norms as analytic of logical statements or judgments and ask whether any bridge principle can be plausibly thought of as analytic. We can test whether a bridge principle is indeed analytic by asking whether anyone who grasps the concept of logical validity or understands the meaning of the term "entails" can sensibly be viewed as having the normative commitments it attributes. I will argue that no bridge principle passes this test, so no principle can be plausibly viewed as analytic of logical statements or judgments.

3.1 Narrow scope

First, consider the class of narrow scope principles, such as (Co+):

> **(Co+)** (P_1, …, $P_n \vDash C$) → if *S* believes P_1, …, P_n then *S* ought to believe *C.*

Now, is it possible for someone who is fully competent with the concept of logical validity to accept an instance of the antecedent while rejecting the relevant instance of the consequent? Using this test, it is clear, for familiar reasons, that (Co+) does not characterise the normative consequences analytically entailed by logical judgments, since philosophers who are fully competent with the concept of logical validity, and who accept that some argument from P_1, …, P_n to *C* is valid, have found sensible grounds to deny that if one be-

lieves $P_1, \ldots, P_n$, then one ought to believe C. For instance, consider Harman's (1986) much discussed "clutter" objection to (Co+): if applied to the rule of Disjunction Introduction ($P \vDash P \vee Q$), (Co+) entails that if one believes P, then one ought to believe P *or* Q for arbitrary Q. Yet, P *or* Q may be a junk belief, of no intrinsic interest, or it may be entirely irrelevant to any of one's practical pursuits. Indeed, (Co+) applied to $P \vDash P \vee Q$ entails an infinite chain of obligations: if one believes P, then one ought to believe P *or* Q, and if one believes P *or* Q, one ought to believe (P *or* Q) *or* R, and if one believes (P *or* Q) *or* R, one ought to believe (P *or* Q *or* R) *or* S, and so on, *ad infinitum*. Moreover, some propositions, such as infinite disjunctions or conjunctions, are so complex that it is not humanly possible to believe them. Yet, if one believes that P, (Co+) entails that one ought to believe P *or* Q even for unbelievable Q. If ought implies can, (Co+) is false.

There are of course various ways to respond to Harman's objection. For instance, one might distinguish between explicit and implicit beliefs, where implicit beliefs are merely dispositions to believe (Field 2009b). (Co+) may not seem to be implausibly demanding if it tells you that if you believe P you must be *disposed* to believe P *or* Q.[6] However, our question here is not so much whether (Co+) is *true*, but whether it is analytic; that is, whether anyone who grasps the concept of logical validity must accept (Co+). And it is clear that (Co+) is not analytic. Harman himself is a case in point: *he* accepts the validity of arguments from P to $P \vee Q$, yet denies that if one believes P, one ought to believe P *or* Q. Since Harman is presumably fully competent with the concept of logical validity, (Co+) is not conceptually necessary.

Another example of a philosopher competent with the concept of logical validity, yet who denies (Co+), is John Broome (2013). One of his many objections to (Co+) is the "bootstrapping worry": given that $P \vDash P$, (Co+) entails that if one *does* believe that P, then one *ought* to believe that P. If one

[6] This response has limitations as well, particularly in the face of propositions that are too complex to be believed. If implicit belief is understood in dispositional terms—as the disposition to have the occurrent belief—then if $P \vee Q$ cannot be occurrently believed (for some unbelievable Q), it cannot be implicitly believed either.

does believe that the number of stars is even, (Co+) entails that one *ought* to believe that the number of stars is even; if one *does* believe that 2+2=5, (Co+) entails that one *ought* to believe that 2+2=5. Yet, one ought to believe no such things, whether or not one already believes them. Once again, this calls the analyticity of (Co+) into question. In this case, Broome is a case in point. *He* accepts that $P \vDash P$, but does not accept that one ought to believe whatever one does believe. Since he is presumably fully competent with the concept of logical validity, (Co+) is not conceptually necessary.

Third, consider the classical principle of Explosion, (EXP) $P \wedge \sim P \vDash Q$, which states that an inconsistent set of premises entails everything. Applied to EXP, (Co+) entails that if you have contradictory beliefs, you ought to believe everything, which is patently absurd. Indeed, paraconsistent logicians have pointed to this consequence to argue that EXP should be rejected (cf. Priest 1979). However, the absurdity of this consequence suggests more strongly still that (Co+) is not conceptually necessary. That is, it is possible for someone to be fully competent with the concept of logical validity, and to accept EXP while quite sensibly denying that if one just happens to have contradictory beliefs, one ought to believe everything. It is implausible that all classical logicians are conceptually confused.[7] All of this suggests that (Co+) does not capture the normative role of the concept of logical validity.

Moreover, the foregoing considerations tell against the analyticity of all narrow scope principles. Just as one might sensibly accept that $P \vDash P$, yet deny that your believing P entails that you ought to believe P, it would be sensible to accept that $P \vDash P$ yet deny that your believing P implies that you are *permitted* to believe P, have a *reason* to believe P, that it is *good* to believe P, or that believing P is what an epistemically virtuous agent would do. Warren (2020, 4.VII), for in-

[7] Priest (1979, 297) charges logicians who accept EXP with a kind of conceptual deficiency. However, it is far more plausible that the concept of logical validity does not have (Co+) as an analytic normative consequence, than that all classical logicians are incompetent with the concept of logical validity. For objections to Priest's argument against classical logic, which assumes the normativity of logic as a premise, see Musgrave (2020) and Steinberger (2016).

stance, suggests that if one accepts the premises of an argument one takes to be valid, this gives one some justification, or some reason for accepting the conclusion. However, this does not seem to give a satisfactory solution to the problem of bootstrapping, since it allows that merely accepting P gives one some justification or reason to accept P, which is implausible, and something Broome would likely deny. The application to EXP is similarly problematic, since it is far from obvious that accepting a contradiction gives one even a modicum of justification, or even a defeasible reason, for believing anything whatsoever. Thus, it would be sensible for a proponent of classical logic to accept the validity of EXP while denying that acceptance of a contradiction provides any justification at all for believing everything.

This goes for bridge principles of negative polarity as well. One might sensibly accept that $P \vDash P$ yet deny that the fact that you believe P entails that you ought not to, are not permitted to, or have no reason to disbelieve P. Each of these narrow scope principles could be sensibly rejected by someone who accepts classical logic without indicating incompetence with or incomplete grasp of the concept of validity.

3.2 Wide scope

Next consider the class of wide scope principles, such as (Wo+):

> **(Wo+)**: (P_1, …, $P_n \vDash C$) → It ought to be the case that: if S believes P_1, …, P_n then S believes C.

Unlike (Co+), (Wo+) seems more promising, since it does not entail that if you *do* believe the premises of a valid argument, then you ought to believe its conclusion. Rather, it entails that you have a conditional obligation to combine believing the premises of a valid argument with believing its conclusion. This wide scope requirement can be satisfied in two ways: either you can satisfy it by both believing the premises of a valid argument and believing its conclusion, or you can satisfy it by not believing one of the premises. For this reason, (Wo+) seems to do better with respect to the bootstrapping worry, since it only entails that you ought to combine believ-

ing P with believing P, which is perhaps redundant, but not obviously false.

However, it is not entirely clear that (Wo+) helps with the clutter objection. Here is one reason why. Suppose that you believe P and accept that $P \vDash P \vee Q$. If (Wo+) is analytic or conceptually necessary, then on pain of incoherence, you must accept that you ought either to not believe anything at all, or to believe all of the logical consequences of your beliefs. Given the implausibility of this normative judgment, it seems that it is possible to sensibly deny it, while still accepting Disjunction Introduction (cf. Broome 2013).

What about the explosion objection? One might think that, on the face of it, (Wo+) deals with it well. (Wo+) applied to EXP can be stated as follows:

> **(Wo+$_{\text{EXP}}$)** $(P \wedge \sim P \vDash Q) \rightarrow$ It ought to be the case that (if one believes both P and $\sim P$, then one believes Q).

(Wo+$_{\text{EXP}}$) does not entail that if you believe both P and $\sim P$, you ought to believe Q. Rather, it only entails that you ought to make sure that you don't *combine* believing both P and $\sim P$ with disbelieving Q. And this might not seem to be so bad, because you can satisfy this normative requirement by either giving up your belief that P or by giving up your belief that $\sim P$. You don't *have* to satisfy it by coming to believe Q.

Nevertheless, (Wo+$_{\text{EXP}}$) is not plausibly analytic, since it too can be sensibly denied without indicating conceptual confusion. First, notice that though believing everything is not the only way to satisfy (Wo+$_{\text{EXP}}$), it is one way to satisfy it. Thus, there is a sense in which (Wo+$_{\text{EXP}}$) assigns a *positive normative status* to your believing everything. Viewed in synchronic terms, it deems a cognitive system that contains a belief in every proposition and its negation to be normatively ideal. Viewed in diachronic terms, if you discover that you have contradictory beliefs, and then form the belief that snakes ride bicycles, (Wo+$_{\text{EXP}}$) applauds your inference: it entails that you have done *something* that you ought to do. Of course, in adding one arbitrary belief, you have not done everything that you ought to do, since given that you have contradictory beliefs, (Wo+$_{\text{EXP}}$) entails that you ought to either give one of them up or come to believe everything, but by coming to form one arbitrary additional belief, you have

come one step closer to believing everything; you have done a *part* of what you ought to do, and thus have done something laudable by the lights of (Wo+$_{EXP}$). This in itself constitutes sensible grounds to deny (Wo+$_{EXP}$).

One might attempt to respond to these worries by appeal to the Law of Non-Contradiction, $\sim(P \wedge \sim P)$ (Field 2009b). A logician who accepts this law will judge that it is never the case that one ought to believe both P and $\sim P$. If this is taken together with EXP, then the two normative principles together entail that the only permissible way to satisfy (Wo+$_{EXP}$) is by ceasing to have contradictory beliefs. However, this response does not address the basic point here. Even if you accept the Law of Non-Contradiction, insofar as you still accept (Wo+$_{EXP}$), you assign *some* positive normative status to believing P, $\sim P$ and Q. And this in itself constitutes sensible grounds for rejecting (Wo+$_{EXP}$).

Moreover, there is a further difficulty with treating (Wo+$_{EXP}$) as analytic that is untouched by the appeal to the Law of Non-Contradiction. The difficulty is this: there are some rules of deontic logic, which would permit one, under certain conditions, to infer that one ought to believe Q, given that one believes both P and $\sim P$. These rules may be controversial, but accepting them seems at least to be compatible with having a full grasp of the concept of logical validity. For instance, Sven Danielsson (2005), who we can presume is competent with the concept of logical validity, put forward the following principle, where O is the deontic operator "ought, " the subscript "i" is an index to a time, X and Y are acts, and N is a modal operator such that NX means that X is inevitable, either because it has actually occurred, or because the option whether to do X is for one reason or the other not open to the agent:

Detachment. $O_i(X \rightarrow Y) \wedge N_i(X) \vDash O_i(Y)$.

If (Wo+$_{EXP}$) captures the normative commitments of someone who accepts EXP, then someone like Danielsson, who also accepts Detachment, is committed to judging that at least in those circumstances in which it is inevitable that one has contradictory beliefs, one ought to believe Q, for arbitrary Q. Moreover, it seems plausible that there *are* circumstances in which it is inevitable that one has contradictory beliefs. For

instance, one might arrive at inconsistent beliefs as a result of complex reasoning in separate contexts, and one might not have noticed the inconsistency because the inconsistent systems of beliefs have not been brought together. If one is not aware of an inconsistency, or perhaps cannot be made aware of it due to the complexity of each belief system, then there is a sense in which eliminating the inconsistent beliefs is not really an option. Or perhaps one discovers that one has inconsistent beliefs but finds that each belief is so well-supported by the evidence that it is difficult to know which one to give up. In such a situation it seems as though having inconsistent beliefs is in a certain sense inevitable, at least for the period of time during which one does not know which belief to give up. In both of these kinds of situations, Detachment together with (Wo+$_{\text{EXP}}$) entail that one ought believe Q, for arbitrary Q—which Danielsson would quite sensibly deny. Thus, it seems to be possible to be fully competent with the concept of logical validity without accepting (Wo+$_{\text{EXP}}$), so (Wo+$_{\text{EXP}}$) is not analytic of the concept of logical validity.

Do similar difficulties arise for wide scope principles involving different normative operators? Consider, for instance, the following alternatives:

(Wp+$_{\text{EXP}}$) $(P \wedge \sim P \vDash Q) \rightarrow$ It is permitted that (if one believes both P and $\sim P$, then one believes Q).

(Wr+$_{\text{EXP}}$) $(P \wedge \sim P \vDash Q) \rightarrow$ There is a reason that (if one believes both P and $\sim P$, then one believes Q).

The foregoing difficulties carry over to these principles too, since both of them assign a positive normative status to simultaneously believing P, $\sim P$ and Q, for arbitrary Q: the first entails that this state is permissible, while the other entails that one has a reason to be in it. Yet, both entailments might sensibly be rejected by someone who accepts EXP.

Wide scope principles with negative polarity, on the other hand, seem to be non-starters. For instance, consider (Wo–$_{\text{EXP}}$):

(Wo–$_{\text{EXP}}$) $(P \wedge \sim P \vDash Q) \rightarrow$ It ought to be the case that (if one believes both P and $\sim P$, then one does not disbelieve Q).

Intuitively, EXP entails that from a contradiction, anything follows. Yet, if we understand disbelieving Q to be equivalent to believing $\sim Q$, (Wo–$_{EXP}$) entails that one way to satisfy (Wo–$_{EXP}$) is to believe P, believe $\sim P$, and not believe $\sim Q$, though $\sim Q$ is just as much a consequence of $P \wedge \sim P$ as Q.

3.3 Distributed

Perhaps distributed norms do better with respect to EXP. For instance, consider (Do+) applied to EXP:

> **(Do+$_{EXP}$)** $(P \wedge \sim P \vDash Q) \rightarrow$ If S ought to believe P, and S ought to believe $\sim P$, then S ought to believe Q.

On the face of it, (Do+$_{EXP}$) seems more plausible than the previous principles, since it entails that you ought to believe Q only if you *ought* to believe both P and $\sim P$. And it might be argued that there are *never* circumstances in which you ought to both believe P and believe $\sim P$. As a consequence, acceptance of this normative principle will never commit you to accepting that you ought to believe anything whatsoever.

However, the assumption that there are never circumstances in which you ought to have contradictory beliefs is questionable. An obvious way to put pressure on it is by appeal to the Preface Paradox (cf. Steinberger 2016). Suppose that Sita has written a book about birds. She has researched it very carefully, and has good evidence for each of the statements that she makes in the book. Let P be the conjunction of these statements. On evidential grounds, it seems that Sita ought to believe P. Yet, Sita is also rightly aware of her own fallibility. Since it is a very long book, she has excellent reason to think that at least one of the statements in it is false. Indeed, if she has very good evidence of her own fallibility, Sita arguably *ought* to think this; she ought to think that $\sim P$. In such a context, acceptance of (Do+$_{EXP}$) entails that Sita ought to believe everything.

It might be objected that this is not the correct account of the Preface Paradox. Perhaps it will be argued that though Sita ought to believe each of the statements in the book, she ought not to believe their conjunction. This is certainly one prominent response to the paradox (cf. Kyburg 1961). However, the question we are considering here does not concern

the best way to resolve the Preface Paradox, but the question of whether (Do+$_{EXP}$) captures the normative commitments one must have in order to accept EXP, with full grasp the concept of logical validity. Moreover, there are logicians who are fully competent with the concept of logical validity, and who accept not only EXP but also Agglomeration ($P, Q \vDash P \wedge Q$). If grasp of the concept of logical validity gives rise to a distributed normative commitment such as (Do+), anyone who grasps the concept of validity and accepts Agglomeration, is committed to the following:

> **(Do+$_{CI}$)** ($P, Q \vDash P \wedge Q$) → If S ought to believe P, and S ought to believe Q, then S ought to believe $P \wedge Q$.

From (Do+$_{CI}$), it follows that anyone who accepts Agglomeration and who grasps the concept of logical validity must judge that Sita ought to believe the conjunction of all the statements in her book, given that she ought to believe each one individually. Thus, someone who accepts both EXP and Agglomeration is committed to saying that Sita ought to believe anything whatsoever, given that she ought to believe both the conjunction of statements in her book, and that at least one of them is false. Yet, this normative claim can be sensibly denied; so (Do+) is not conceptually necessary.

Once again, the same line of reasoning holds for all of the other distributed principles. Consider, for instance, the principle that states that if you have reason to believe the premises of a valid argument, you have reason to accept the conclusion, which Steinberger suggests may help with the preface paradox (Steinberger 2019, 25):

> **(Dr+$_{CI}$)** ($P, Q \vDash P \wedge Q$) → If S has reason to believe both P and Q, then S has reason to believe $P \wedge Q$.

However, while this seems to be plausible as a normative consequence of Agglomeration, even in the face of the preface paradox, it does not obviously capture the analytic consequences of accepting EXP:

> **(Dr+$_{EXP}$)** ($P \wedge \sim P \vDash Q$) → If S has reason to believe P, and reason to believe $\sim P$, then S has reason to believe Q.

A classical logician who accepts EXP can coherently do so while quite sensibly rejecting the normative consequences as

postulated by (Dr+). If one has a mixed bag of evidence, some of which supports P, and some of which supports $\sim P$, one arguably has reason to believe P, and reason to believe $\sim P$, yet no reason to believe Q, for arbitrary Q. Similarly, a classical logician can coherently accept the validity of EXP while denying that if one is permitted to believe P, and permitted to believe $\sim P$, then one is permitted to believe Q, or that it is good to believe Q, or that it would be virtuous to believe Q, and so forth.

3.4 Credence

The foregoing principles involved full belief. But it may be that the solution to the foregoing difficulties lies in formulating the normative principles in terms of degrees of belief, or credences. For instance, Field's view (at least in one of its formulations) is that the normative commitments that come along with judging an argument to be valid involves the commitment to a policy constraining on one's degrees of belief as follows:

> (VP_a): To regard the argument from P_1, ..., P_n to Q as valid is to accept a constraint on degrees of belief: one that prohibits having degrees of belief where $Cr(Q)$ is less than $\Sigma\, Cr(P_i) - n + 1$; i.e., where $Dis(Q) > \Sigma_i\, Dis(P_i)$).

Here $Dis(P) = 1 - Cr(P)$, and can be written as "your disbelief in P." Field's principle, simply put, says that if you regard an argument as valid, you should not be less certain of the conclusion than you are of the premises taken together. Note that Field's principle does not contain any deontic operators, and does not make it clear whether the implicit deontic operators should be assumed to take wide scope, narrow scope, or to be distributed over the conditional. Let us suppose that he endorses the distributed, ought principle (Do+), which when stated in Field's terms can be understood as follows:

> **(Do+$_{\text{FIELD}}$)** ($P_1, \ldots, P_n \vDash C$) $\rightarrow$ if $\Sigma_i Dis(P_i)$ ought to be n, $Dis(C)$ ought to be $\leq n$.

In other words, if someone who is competent with the concept of logical validity judges that an argument is valid, she must judge that one's disbelief in the conclusion ought not to

exceed the disbelief one ought to have in the premises. Does framing the principle in terms of credence rather than full belief help to resolve the difficulties posed by the Preface Paradox?

It might seem to. After all, Sita's evidence for any one of the statements in her book, though good, falls short of warranting certainty. And when these statements are conjoined, the uncertainties add up, to the point where Sita's rational credence in the conjunction may wind up being rather low. If the book is long, and contains many statements, then the credence Sita ought to have in the conjunction may be low enough not to count as full belief. In this context, it is not the case that Sita ought to believe the conjunction of statements in her book, and hence, even granting assumptions about human fallibility, it is not the case that Sita ought to have contradictory beliefs.

However, this response to the puzzle, though plausible, is not immune to counterexamples. Imagine that instead of writing a book about birds, Sita chose to write a book of mathematics. As it happens, every statement in her book is a necessary truth, so the credence she ought to have in each statement in her book is 1. Yet, she has excellent evidence of her own fallibility—though an accomplished mathematician, she has still caught herself making mistakes from time to time—so she has reason to believe that at least one of the statements in her book is false. In this case, the lowest credence that Sita is permitted to have in the conjunction of all the statements in her book is 1, and this must qualify as full belief. If this is in principle possible, then it is at least in principle possible to construct a case in which Sita ought to believe both P and $\sim P$. This gives us good reason to deny (Do+$_{\text{FIELD}}$). As in previous cases, this point generalizes to distributed principles involving alternative normative operators.

4. Concluding remarks

I have considered several proposals regarding the normative consequences of logical statements or judgments. Yet, none of those I have considered have a plausible claim to be analytic, since it seems possible for someone who is competent with the concept of logical validity to judge that an argument form

is valid, while rejecting the normative consequences that are purported to follow from accepting this. It is possible that there are alternatives that I have not considered. I cannot claim to have been exhaustive. However, given the range of principles I have considered, we seem to have good reason to think that NL is false, and that logical statements and judgments do not have normative consequences analytically.

One potential response to this line of objection to NL is to point out that it implicitly assumes that the normative consequences of logical statements or judgments must be systematic across all logical principles that one might take to be valid. Justification for this assumption derives from the fact that the normative consequences of logical judgments plausibly derive from the logical concepts they contain, such as the concept of logical validity or entailment. If that is so, then one should expect that the normative consequences of validity judgments remain constant, whether one thinks that EXP or Agglomeration is valid. However, a logical pluralist might be inclined to resist this assumption, and argue that the normative consequences of validity judgments vary from person to person, and that the contents or truth values of validity statements vary from context to context. Such a response would make communication and disagreement about logic well-nigh impossible, since it would imply that both the descriptive content and the normative content of logical statements would vary, leaving no shared language in which to communicate (Hattiangadi 2018b). Thus, the response comes at a significant cost. On balance, then, I conclude that there seems to be good reason to reject the view that logic is normative.

Stockholm University

References

Ayer, A. J. (1946). *Language, Truth and Logic*, 2nd ed. London: Victor Gollancz.

Ayer, A. J., C. H. Whitely, and M. Black (1936). "Truth by Convention: A Symposium." *Analysis* 4(2–3), 17–32.

Beall, J. C. and Greg Restall (2006). *Logical Pluralism*. Oxford: Oxford University Press.

Broome, John (2013). *Rationality Through Reasoning*. Oxford: Wiley-Blackwell.

Carnap, Rudolf (1937/2001). *The Logical Syntax of Language*. London: Routledge.

Caret, Colin R. (2016). "The Collapse of Logical Pluralism has been Greatly Exaggerated." *Erkenntnis* 82(4), 739–760.

Danielsson, Sven (2005). "Taking Ross's Paradox Seriously: A Note on the Original Problems of Deontic Logic." *Theoria* 71(1), 20–28.

Dutilh-Novaes, Catarina (2015). "A Dialogical, Multi-agent Account of the Normativity of Logic." *Dialectica* 69, 587–609.

Field, Hartry (2009a). "What is the Normative Role of Logic?" *Proceedings of the Aristotelian Society* 83, 252–68.

Field, Hartry (2009b). "Pluralism in Logic." *Review of Symbolic Logic* 2(2), 342–359.

Field, Hartry (2009c). "Epistemology without Metaphysics." *Philosophical Studies* 143, 249–290.

Field, Hartry (2015). "What is Logical Validity?" In Colin R. Caret and Ole T. Hjortland (eds.), *Foundations of Logical Consequence*. Oxford: Oxford University Press.

Field, Hartry (2022). "Conventionalism about Mathematics and Logic." *Noûs*. https://doi.org/10.1111/nous.12428

Frege, Gottlob (1897/1997). "Logic." In Michael Beaney (ed.), *The Frege Reader*. Oxford: Blackwell, 227–250.

Frege, Gottlob (1918/1997). "Thought." In Michael Beaney (ed.), *The Frege Reader*. Oxford: Blackwell, 325–345.

Glüer, Kathrin, and Åsa Wikforss (2009). "Against Content Normativity." *Mind* 118, 31–70.

Harman, Gilbert (1986). *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Hattiangadi, Anandi (2003). "Making it Implicit: Brandom on Rule Following." *Philosophy and Phenomenological Research* 66, 419–431.

Hattiangadi, Anandi (2007). *Oughts and Thoughts: Rule-Following and the Normativity of Content*. Oxford: Oxford University Press.

Hattiangadi, Anandi (2017). "The Normativity of Meaning." In Bob Hale, Crispin Wright, and Alexander Miller (eds.), *A Companion to the Philosophy of Language*, 2nd edition. Oxford: Wiley Blackwell, 649–669.

Hattiangadi, Anandi (2018a). "Moral Supervenience." *Canadian Journal of Philosophy* 48(3–4), 592–615.

Hattiangadi, Anandi (2018b). "Logical Disagreement." In Conor McHugh and Daniel Whiting (eds.), *Meta-Epistemology*. Oxford: Oxford University Press, 88–106.

Hattiangadi, Anandi (2023). "Logical Conventionalism and the Adoption Problem." *Aristotelian Society Supplementary Volume* 97(1), 47–81.

Kant, Immanuel (1781/1787/1988). *Kritik Der Reinen Vernunft* (Critique of Pure Reason). Hamburg: Meiner.

Keefe, Rosanna (2014). "What Logical Pluralism Cannot Be." *Synthese* 191, 1375–1390.

Kripke, Saul A. (forthcoming). "The Question of Logic." *Mind*.

Kyburg, Henry (1961). *Probability and the Logic of Rational Belief*. Middletown: Wesleyan University Press.

MacFarlane, John (2002). "Frege, Kant, and the Logic in Logicism." *The Philosophical Review* 111, 25–65.

MacFarlane, John (2004). "In What Sense (if any) is Logic Normative for Thought?" Unpublished manuscript.

MacFarlane, John (2017). "Is Logic a Normative Discipline?" Presentation at the conference on the Normativity of Logic, University of Bergen, June 14, 2017.

Musgrave, Alan (1972). "George Boole and Psychologism." *Scientia* 107, 593–608.

Musgrave, Alan (2020). "Against Paraconsistentism." In Wenceslao J. Gonzalez (ed.), *New Approaches to Scientific Realism*. (Epistemic Studie 42). Berlin & Boston: De Gruyter, 133–144.

Pettigrew, Richard (2017). "Epistemic Utility and the Normativity of Logic." *Logos and Episteme* 8(4), 455–492.

Pigden, Charles, and Elizabeth Olsen, "The Normativity of Logic and No-Ought-From-Is." Unpublished manuscript.

Priest, Graham (1979). "Two Dogmas of Quineanism." *Philosophical Quarterly* 117, 289–301.

Railton, Peter (2000). "A Priori Rules: Wittgenstein on the Normativity of Logic." In Paul Boghossian and Christopher Peacocke (eds.), *New Essays on the a Priori*. Oxford: Oxford University Press, 170–96.

Read, Stephen (2006). "Monism: The One True Logic." In D. DeVidi and T. Kenyon (eds.), *A Logical Approach to Philosophy: Essays in Honour of Graham Solomon*. Berlin: Springer, 193–209.

Russell, Gillian (2017). "Logic Isn't Normative." *Inquiry*. DOI: 10.1080/0020174X.2017.1372305.

Sainsbury, Mark (2002). "Which Logic Should We Think With?" In A. O'Hear (ed.), *Logic, Thought, and Language*. Cambridge: Cambridge University Press, 1–17.

Steinberger, Florian (2016). "Explosion and the Normativity of Logic." *Mind* 125(498), 385–419.

Steinberger, Florian (2017a). "The Normative Status of Logic." In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 edition) <https://plato.stanford.edu/archives/spr2017/entries/logic-normative/>.

Steinberger, Florian (2017b). "Frege and Carnap on the Normativity of Logic." *Synthese* 194(1), 143–162.

Steinberger, Florian (2019). "Three Ways in Which Logic Might Be Normative." *Journal of Philosophy* 116(1), 5–31.

Warren, Jared. (2020). *Shadows of Syntax: Revitalizing Logical and Mathematical Conventionalism*. New York: Oxford University Press.

Williamson, Timothy (2007). *The Philosophy of Philosophy*. Oxford: Blackwell.

Woods, Jack (2023). "A Sketchy Logical Conventionalism." *Aristotelian Supplementary Volume*, 97(1), 29–46.

Constitutive Rules and Internal Criticism of Assertion

JAAKKO REINIKAINEN

1. Introduction

Timothy Williamson famously proposed that the speech act of assertion is constituted by the knowledge rule:

> *K-rule.* One must: assert that *p* only if one knows that *p*. (Williamson 1996, 494)

The K-rule is "constitutive" in the sense that no speech act can count as an assertion without being subject to this rule; the K-rule is necessarily (and uniquely) in force for assertion. Nonetheless, it is perfectly possible and potentially even common for assertions to fail to live up to the K-rule without ceasing to be assertions. For example, a liar who knows that not-*p* may still assert that *p*.

If there is a K-rule (or another comparable epistemic rule) constitutive of assertion, then this has important consequences for the criticism of assertions. To put it simply, an assertion can be criticized on contingent or "external" grounds, e.g., for being impolite, or then it can be criticized on necessary "internal" grounds, e.g., for being unjustified or false. Compare the case to a game, for example chess, where a move is subject to internal criticism if it violates the rules of chess. The reason or justification for the criticism is then internal to chess in the sense that it is impossible to play chess without being governed by the rules to which the criticism appeals, simply because the rules are constitutive of chess.

However, there is an important difference. The rules of chess are explicitly defined, but there is no rule book of the game of assertion to which speakers can appeal. If there are constitutive rules of assertion in Williamson's sense, these must in the first instance be grasped implicitly (Williamson 1996, 492). The main evidence for identifying the constitutive

rules of assertion is left to intuitions, derived either from real observations or thought experiments, as to how we would evaluate assertions in various situations.

The problem developed in this paper concerns the question of how we are to discern not which rules are intuitively really constitutive of assertion, but whether intuitions can differentiate between the internal and external reasons for criticism in the case of assertion to begin with. The main contention is thus broadly methodological in nature: can intuitions provide evidence for what the normative source for criticizing an assertion is? In particular, can our intuitions distinguish between reasons appealing to the "normative," rule-constituted nature of assertion as opposed to more generic social or moral norms of conduct?

There are two reasons why this question matters. The first is that intuitions about the source of criticizing assertions have been used as evidence for the constitutive rule account of assertion. The second is that intuitions have also been used as evidence against competing accounts (Goldberg 2015; MacFarlane 2011). In particular, it is claimed the so-called attitudinal account of assertion (Bach and Harnish 1979) cannot explain certain essentially normative features of assertion which rely on the idea of internal criticism, namely authorization and retraction. My positive claim here is that the attitudinal account can in fact respond to the evidence from intuitions because it is not clear that intuitions can decide whether there is a distinction between internal and external criticism in the case of assertion. At the very least, there are complications involved that have not been considered before. Finally, I show that arguments similar to the case of assertion have been raised in the case of institutional roles, e.g., being a professor (Roversi 2021), which are also claimed to include implicit rules that allow internal forms of criticism. I apply a similar line of criticism against Roversi's proposal.

Before moving on to the arguments, it is good to emphasize that since the paper's angle is methodological, it does not aim to attack the very idea of the constitutive rules account. For all that is said here, there could be constitutive rules of assertion or of being a professor. The question is what can be counted as evidence in favor of deciding the matter.

2. Internal and external criticism

Starting with the proposal that assertion is constituted by some internal, epistemic norm, in this section I look into the debate on how this premise can be used to argue against competing accounts of assertion. But to begin with, it is useful to say a few more words about the distinction between internal and external criticism in this context. I will also present the outline of the attitudinal account of assertion which stands in contrast to the rule-constituted account. Finally, I present the objections to the attitudinal account by Goldberg (2015) and MacFarlane (2011), which draw from the possibility of internal criticism and assertion being constituted by rules.

As already mentioned, the distinction concerns the grounds of criticism, or the justifying reasons one can have for criticizing a given assertion. In both, the case of internal and external criticism relevant here; the form of justification is to appeal to a rule or norm (I use these terms interchangeably) which the assertion violates. In the case of chess, this is an easy distinction to make because chess rules are relatively (a) clearly articulated and (b) easy to monitor. In most cases, we can confidently say whether a given action by a chess player is to be criticized on the grounds that it violated the rules of chess or because it violated some more generic social rule, e.g., being unsportsmanlike (naturally it could be both). Williamson's original idea was that something similar is true of assertion, though here the distinction is not explicit but must be discovered by philosophical argumentation backed up by intuitions.

In order to question the evidence from intuitions regarding the source of criticizing assertions, it is useful to have as a contrast an account of assertion which does not entail the possibility of internal criticism of assertions. Here that role falls on the attitudinal account as developed by Bach and Harnish (1979). Before outlining the account, I want to emphasize that my main aim is not to provide novel arguments in favor of the attitudinal account, but rather to show that it is on equal footing with the constitutive rules account when it comes to evaluating evidence from intuitions. In order to show this, however, the attitudinal account may need to be adjusted somewhat, as I shall do below. My new suggestion is that the attitudinal account can incorporate the idea that assertion is a device for expressing knowledge and not mere-

ly beliefs without taking on the idea that assertion as such is constituted by some epistemic norm.

Briefly, the attitudinal account as developed by Bach and Harnish (1979) claims that assertion is defined by two conditions:

> In uttering e, *S* asserts that *P* if *S* expresses:
>
> i. the belief that *P*, and
>
> ii. the intention that *H* believe that *p*. (Bach and Harnish 1979, 42)

In this (simplified) picture, to make assertions is to express beliefs with the "R-intention" of giving the audience a reason to ascribe the belief that *p* to the speaker while also purporting to make *H* also believe that *p*. To "R-intend" means to cause beliefs in the audience by the way of their recognition of this very intention, loosely on the speech act model of Grice (1957). The speaker can thus (purport to) cause beliefs in others by asserting claims, which in large part explains why people make assertions at all.

The important point here is that the attitudinal account of assertion does not involve the possibility of internal criticism of assertions because assertion thus described is not a normatively *special* way to intentionally cause beliefs in others. Of course, a speaker will usually be (held) responsible for her assertions, and she may be criticized if her assertion turns out to be, e.g., false, unjustified, or impolite. Furthermore, criticism according to which the assertion was unjustified may be in many ways more pertinent (in the context) than criticism according to which it was impolite. Yet the source of the criticism or its pertinence is not in assertion's internal rules but in more general social or moral rules and norms, which govern all actions indiscriminately.

Some authors think the normatively indiscriminate treatment is wrong. Goldberg for one thinks that the attitudinal account is wrong to miss the internal form of criticism for assertions. He provides the following example:

> Compare: I may know that these cookies are for Ralph, and even so I may place them in a spot where I know you will encounter them, intending that you eat them (by way of your recognizing my intention). Still, if you do, it is no excuse to say that I authorized you to eat them—I did no such thing! To *tempt* a person to φ (by doing something with the intention that they φ by way of

> their recognizing this intention) is not the same as *authorizing* her to φ. In short, even if the asserter intends the hearer to form the belief in question, intending is one thing, *authorizing* is another, and it would seem that the attitudinal view has no basis for moving from the former to the latter. What is missing here, and what the attitudinal view seems to fail to deliver, is a sense that S *ought not* to have asserted as she did—and that the reason she ought not to have done so is that in so asserting she *authorized* H to believe as he did. (Goldberg 2015, 14)

In summary, Goldberg's thought is this: Imagine yourself in the position of the oblivious cookie-eater. Were Ralph to ask you why you ate his cookies, your justification might be something like "Because person *S* gave me a reason to believe the cookies were meant for me." In return, *S* would then say that he merely tempted you to believe that, and thus is not responsible for your mistaken belief. Had *S asserted* that the cookies are meant for you, that would be another thing, for then *S* would have *authorized* your belief as opposed to merely intentionally causing it. (Of course, both the attitudinal account and the rule-constituted account can agree that there is a distinction between intentional misleading and outright lying, yet this is not what the cookie example is meant to showcase.)

Before raising objections to Goldberg's criticism, I want to compare it to a parallel objection that has been made against the attitudinal account by MacFarlane (2011).[1] According to MacFarlane, the attitudinal account cannot make sense of the possibility that an assertion can be retracted, which means "rendering [the assertion] 'null and void'" (2011, 83). His idea seems to be that, while it is arguably impossible to "undo" the causal or perlocutionary effects of a token assertion, it should be possible in principle to undo a token assertion's illocutionary effects by retracting it. Since according to the attitudinal account of assertion, the main illocutionary effect of an assertion is to R-intend the audience to ascribe the belief that *p* to the speaker, the account cannot understand retraction literally as an undoing of the intention but rather as an "unexpression":

---

[1] MacFarlane (2011) does not as such defend the Williamsonian account of constitutive rules, but his critical arguments do fit well with that account. In any case, Goldberg (2015, 14–15) uses retraction as a rule-constituted feature of assertion which the attitudinal account fails to explain.

> In uttering e, *S* retracts the claim that *P* if *S* expresses:
>
> i. that he no longer believes that *P*, contrary to what he previously indicated he believed, and
>
> ii. the intention that *H* not believe that *P*. (Bach and Hamish 1979, 43)

However, according to MacFarlane this will not do:

> One can, without any insincerity, retract an assertion of something one still believes. One might do this, for example, because one realizes one can't adequately defend the claim, or because one doesn't want others relying on it. Indeed, it is possible to retract the assertion while avowing the belief: "I retract that, as I can't defend it. But I still believe it." This does not seem insincere in the way that "I assert that *p*, but I don't believe it" does. So it does not seem right that retraction expresses lack of belief. Nor does it express an intention that one's audience not believe what was asserted—one may be quite happy to let them continue to believe this, if they have their own independent grounds. (MacFarlane 2011, 83)

The reason why this objection is parallel to the point raised by Goldberg about "authorization" is that if we think of assertion as coming with internal norms, there is a ready way to think how an assertion can be retracted in the literal sense in which the word is used, e.g., in publishing and law. Retracting an assertion would then be like retracting a move in a game. While this does not undo the causal, perlocutionary effects of the move, it will return the game to a state prior to the move. Analogously, MacFarlane suggests that in asserting "I retract that *p*, but I still believe it," one gives up the responsibility of defending the assertion, thus its illocutionary effect, while letting the audience think that the speaker still believes that *p*, and that (possibly) they should too.

3. Responding to objections to the attitudinal account

In this section, I look into how the attitudinal account can respond to the two objections raised above, beginning with Goldberg.

Goldberg's starting point is to contrast two ways in which the speaker can intentionally spread beliefs to his audience, then to claim that assertion plays this role in a normatively

special way. The cookie example brings out the intended contrast well: It is a different thing to intentionally cause someone to believe that *p* with the knowledge that this was the speaker's intention than it is to authorize the audience to believe that *p*. The difference comes down to the justification for criticizing the speaker. According to the attitudinal account, an assertion can be criticized for being misleading or outright lying on generic moral or social grounds, whereas according to Goldberg, following Williamson, there is an additional normative source involved, namely the internal, constitutive epistemic norms of assertion. The problem I want to raise for Goldberg (and Williamson) is not about this distinction as such but the methods for showing that it is a distinction which we can identify in our ordinary practice of making assertions, as opposed to a theoretical postulate. So, how is it to be settled on what grounds it is justified to criticize an epistemically incorrect assertion?

Terms like "authorize" presume, in their literal (current) meaning, an institutional background with explicit rules, roles, and positions for subjects.[2] Since the ability to make assertions presumably is independent of the existence of any official institutions (no one is officially granted the license to make assertions *per se*), the sense in which an assertion purports to "authorize" the audience to believe that *p* cannot rest on its literal meaning; the meaning has to be either metaphorical or technical. However, Goldberg (2015) nowhere defines what he means by "authorize" as a technical term, and as a metaphor it is hardly helpful in argumentation.

As a helpful reviewer pointed out to me, it could be that Goldberg has in mind the sense of "authorize" which is apparent, e.g., in making a promise, which is arguably an ability independent of official institutional settings. So, asserting that *p* would be like making a promise that *p* is true (or justified

[2] Of the current meanings of "authorize," two are worth noting here. The first is the meaning in which official roles, duties, powers, etc. are given in the legal sense. The second is to justify actions in general. Goldberg cannot have in mind the second meaning because this sense is agreeable to the attitudinal account: In making an assertion that *p*, the speaker aims to give the audience *a good* reason to believe that *p*. In case the reason is not in fact good, the speaker has generically misled the audience ("authorize, v." OED Online. June 2022. Oxford University Press. https://www.oed.com/view/Entry/13352?redirectedFrom=authorise (accessed August 23, 2022)).

etc.), or swearing that it is. If true, I agree that this would show asserting to be something different than to R-intend the audience to form the belief that *p*, just as it is different to R-intend the audience to believe that the speaker will do something than it is to promise to do something.

However, the question is how we are to make this distinction at the level of intuitions about imaginary or real cases of assertion, not whether the distinction is clear in the abstract. It is clear that the attitudinal account has no difficulties in explaining why we criticize assertions on epistemic grounds: The reason is simply that we care to have true, justified beliefs expressed to us. In the cookie example, *S* is responsible for misleading the speaker as to whom the cookies are meant, or if *S* expressed himself by way of an assertion, for lying. According to Goldberg, *S* is responsible in the additional sense for having violated a constitutive epistemic rule of assertion (whatever that exactly is). But since we cannot appeal to the explicit rule book of assertion as we can in chess, how are we to discern whether the criticism really is external or internal in kind? Moreover, assuming that there are independent social norms and moral norms against misleading and lying, why should we expect assertion to be additionally governed by an internal norm to this very effect?

We can press this question and its point further by adjusting the attitudinal account somewhat. Suppose that in asserting that *p*, the speaker does not merely R-intend the audience to believe that *p*, but makes a knowledge claim that *p*, i.e., presents himself as knowing that *p*. This is possible, let us suppose, because it is a function of assertions to express knowledge. Nonetheless, I argue, the attitudinal account could still hold that there is no constitutive norm of assertion, and hence no distinction between internal and external criticism of incorrect assertions.

First of all, the intended function here is teleological in kind. It could be that our practice of making assertions developed for the purpose of expressing knowledge, i.e., that this function causally explains why we have this practice, akin to how the ability of the heart to pump blood explains why there are hearts. Similarly, the designed function of binoculars to see into the distance is what explains why we have them. The ultimate explanation for these things comes down to the fact that knowledge matters to us, as does seeing far. The important point in regard to the distinction between external and internal criticism of assertion is that if assertion has the teleological function to express knowledge, that is com-

patible with its lacking internal, constitutive epistemic norms. The reasons to criticize faulty assertions would then be broadly the same as the reasons to criticize faulty binoculars, namely that they do not serve their designed function. This is starkly different from internal criticism in the case of chess: there is no functional fault present in moving a rook diagonally because the rook did not develop to move only linearly. Rather, it was stipulated to be governed by this rule.

I move on to MacFarlane's objection that the attitudinal account cannot explain the possibility of retracting an assertion. To be sure, there is a clear sense in which the speaker can retract an assertion that *p* while continuing to believe that *p* that is common to, e.g., law and publishing. But is the same sense so evidently available in informal contexts, so that there is always—or ever—a clear distinction between saying "I didn't mean that" and "I take that back" as MacFarlane claims? As always, the data gathered from actual speech acts is as messy as it gets, but one should think that in quite many contexts saying that "I retract that *p*, but I still believe that *p*," is bound to raise a few eyebrows, for saying "I believe that *p*" is a conventional (if roundabout) way to assert that *p*, or at any rate present that *p* as true. So, there are bound to be at least some contexts where the utterance is backhanded and thus insincere, contrary to MacFarlane.

In any case, suppose there are some informal contexts where the possibility of retracting an assertion while holding onto the belief as well as the R-intention for the audience to believe it is clearly available. Returning to the cookie example, assume that I cannot justify, beyond my testimony, the claim that Ralph misled me to eat the cookies meant for you.[3] Then I might "drop" the claim while continuing to believe this, and also R-intending you to believe it. To express this kind of *partial* retraction, I might say something like "Forget about it, let's move on." Now, this would not be a full retraction because I would not give you a reason to disbelieve my original claim that it was Ralph who misled me to eat your cookies; I simply stop treating the justification of my original claim as pertinent.

The question to MacFarlane becomes this: How credible is it that I could *fully* retract my claim that Ralph made me eat your cookies while also continuing to R-intend you to believe this? If retracting a claim primarily means, following the rule-

[3] For the sake of convenience, I switched the roles in the example.

constituted account, to give up one's epistemic credentials to it, in the full sense this should imply that one cancels the reasons for one's original claim and not merely stops actively defending it. In the cookie example, this would mean changing my original testimony that it was Ralph who misled me by sincerely saying, e.g., "I was wrong about Ralph, the fault was my own after all." But is it really coherent to both sincerely present reasons to cancel the justifications for the original claim (i.e., to fully retract an assertion as opposed to merely "dropping" it) *and* continue to R-intend the audience to believe the original claim? At least in the context of the cookie example, this seems barely coherent: I would both have to defend the (sincere) claim that it wasn't Ralph's fault that I ate your cookies while also R-intending you to believe it was Ralph's fault.

The reason for why it is harder to pry apart the epistemic credentials for assertion and R-intentions than MacFarlane appears to think, I contend, is that in most cases it is precisely the (implicit) epistemic credentials by the way of which we R-intend the audience to form the belief expressed by our assertions. This is compatible with my earlier suggestion that the attitudinal account could be adjusted so that it is the teleological function of assertions to express knowledge. This idea is natural enough: If I want you to believe that *p*, a good way to do this is surely to present *p* as knowledge. If I want to see far, I should use binoculars. But that does not imply that there is a constitutive norm for binoculars such that they should enable one to see far.

4. On being a professor

In this section, I will consider the intuitive evidence for the distinction between internal and external criticism in the context of an institutional role, e.g., being a professor, as defended by Corrado Roversi (2021). Although the topic is different, the focus of my main argument is the same, namely, to question the intuitive evidence for the possibility of internal criticism enabled by constitutive rules.

Roversi builds his case on a thought experiment centered on one Mr. Colasanti, a student in legal philosophy who comes to his professor (Roversi himself) to get help passing his exam. After hearing him out, Roversi clearly perceives that he does not have the time required to ensure that Mr. Colasanti will pass the exam; moreover, Roversi in his position as a professor is not obliged by the explicit rules of his

institution to go beyond the extra mile to help him. Yet Mr. Colasanti is not satisfied with this reasoning; he goes on to demand the extra help precisely by appealing to Roversi's position as a professor despite knowing full well that the official rules do not mandate him to do that. Roversi summarizes his view of the situation as follows:

> I take this retort by Mr. Colasanti to be perfectly meaningful and genuine, something I must reply to with good arguments. His point is that independently of the formal rules set forth by the university, my being a professor requires me to take his situation into account and do my best to improve his understanding of the subject matter. This is what being a professor means, he is implicitly arguing: it means getting students to understand what is being taught. I insist on my formal duties with him, but for the rest of the day I keep mulling over whether there is something I could do. (Roversi 2021, 14355)

Roversi's claim is that two prominent accounts that seek to explain institutional reality without appeal to constitutive rules, namely Epstein's (2015) grounding approach and the approach of Hindriks and Guala (2015), cannot make sense of Mr. Colasanti's reaction as "meaningful and genuine." The reason is that in making the plea, Mr. Colasanti draws his justification from the *ratio* of being a professor, and that the *ratio* can only be understood by appealing to the constitutive rules of being a professor. For the sake of space, I shall only discuss the case from the point of view of Hindriks and Guala's regulative rules account, which at any rate seems to be the better contrast for my purposes. As in the case of assertion, my main aim is not to provide new arguments in favor of Hindriks and Guala, but merely to argue that the evidence from intuitions to which Roversi appeals can be explained from their perspective as well.

The crucial pivot of Roversi's argument is that the source of justification for Mr. Colasanti's appeal is the position of professorship itself understood as distinct from the regulative rules that define it. Without this normative support, Mr. Colasanti's appeal would either be ingenuine (unjustified and misguided) or then its justifications would have to be grounded in more generic normative sources like compassion. This raises an immediate problem: how are we to tell that Mr. Colasanti's appeal *is* meaningful in the relevant sense? Note that to answer this, it does not suffice to know his motivations for making the plea, as only the source for the

plea's justification is relevant. But insofar as the plea is not justified by the explicit regulative rules of professorship, the claim that it is actually meaningful in the "internal point of view's" sense is *prima facie* no more justified than the negative claim.

Moreover, it is interesting to note that in his more detailed analysis of Mr. Colasanti's plea, Roversi actually says that the plea is not meant to claim that professorship is defined by obligations that go beyond what is explicitly stated in the regulative rules, but rather that these regulative rules *should* be changed in view of the position's *ratio*. He goes on to derive a kind of *reductio* argument from this observation:

> On the regulative-rules account, I could simply reply to [Mr. Colasanti] that what he is saying is meaningless, because the very meaning of the term *professor* is a composition of rules, none of which requires me to do what he is asking. But his reaction is *not* meaningless. His argument is precisely that, even recognizing that there is no rule requiring me to support him beyond class time and office hours, a rule of this kind should be added to the list and be made explicit, given the overall *ratio* of the institutional role "being a professor." But this entails that the meaning of *professor* is not simply a set of conditional regulative rules. To state the point more directly: if one can always build a meaningful argument about changing or adding further rules connected with a status in view of that status's purpose or underlying rationale, the concept of that status cannot simply be reduced to the regulative rules that are connected with it. There is at least one other element of meaning apart from the rules, and this element is the overall rationale behind the connection between conditions and normative consequences—the purpose the institution is built for, one might say. If this further element were not part of the picture, any connection, any arbitrary set of rules could do. (Roversi 2021, 14363)

This paragraph appears to contain a slide in the meaning of "meaning of professor." On the one hand, the meaning of being a professor is given by the explicit, official regulative rules that define the position. On the other hand, "the meaning of being a professor" refers to the *ratio* of being a professor, or the purpose that the role is supposed to play in an institution. As such, the regulative rules account should have no problem in recognizing both senses of "the meaning of being a professor" as legitimate so long as they are not con-

flated. Roversi's argument in contrast presumes that the two meanings must come together, or that the official meaning must somehow contain the *ratio*—otherwise "any arbitrary set of rules could do."

But this seems confused. Of course, if an institutional role has no *ratio* or purpose, then one set of rules defining it will not be better than any other (formal considerations notwithstanding). As it happens, most institutions that are defined by regulative rules have been founded for a purpose which those rules reflect, for better or worse. So, the debate is not about whether the institutions and roles within them have *ratios* or not, but whether this *ratio* is included in the meaning that defines the role. Insofar as the meaning of Mr. Colasanti's plea is that the current official meaning of professorship should be changed in view of its *ratio*, there is nothing that contradicts the account according to which the current meaning of being a professor just is given by the currently official regulative rules; in fact, this interpretation of the plea affirms the regulative rules reading.

The only way in which Mr. Colasanti's case would be problematic for the regulative rules account would be if his plea meant (and was correct to mean) that being a professor included obligations going beyond those defined in the official regulative rules. In that case, being a professor would of course be defined by more than a set of official regulative rules. What would then show that the plea is justified in this sense, i.e., that Roversi *is* (and not merely should be) obliged to help his student out beyond official regulations, not merely due to generic normative sources like compassion but because of the *ratio* of being a professor? As far as I can see, Roversi does not answer this crucial question in the paper.

Insofar as *ratio* is not included in the definition of what it means (in the sense of rules) to be a professor, there is no distinct source of normativity which the regulative rule account would miss. This does not imply that an account that relies predominantly on regulative rules as opposed to constitutive rules would be incapable of accounting for the "internal point of view" on institutions and their roles, for all that the internal point of view requires is deliberation about whether the rules serve their purpose and whether they should be changed. But it is perfectly possible and unproblematic to deliberate a change of rules in view of an institution's overall purpose without presuming that the institution (or a role within it) is defined by constitutive rules.

If that is right, does it follow that there is nothing more to the existence of an institutional entity than the set of regulative rules defining it? Roversi appears to think so, which he takes to be another point against the regulative rules account:

> Ownership can have different rules in different legal systems, yet the institution is taken to be the same across these systems and to be commensurable because the different rules serve a similar *ratio*, namely, to make it possible for legal persons to have something at their exclusive disposal. If the constitutive rules of property in a legal system were simply regulative, the institutions of property in different systems could not be recognized as structurally modified instances of the same institution but would have to be considered altogether different entities. (Roversi 2021, 14366)

Again, the argument here, in my view, pivots on a slide in meaning. On the one hand, two different juridical property systems ascribing different sets of regulative rules for "property" will thereby ascribe different meanings to what it is to be property. On the other hand, the two systems might resemble each other a great deal in other respects save what is literally printed in codices; they might share a historical origin, several social functions, many ritualistic practices, etc. So, the systems are different yet similar at the same time. The question is, if we remove all the legal regulative rules, is there anything left that can be called "the same" institution, namely property? All things being equal, the answer must be yes: what remains are, e.g., the history and social functions of the property institution. So, the regulative rules account is compatible with institutions being something more than their regulative rules, only in a different sense. We can therefore recognize two different regulative rule legal systems as different developments of the "same" institution without thereby assuming that the institution must be defined by underlying constitutive rules.

5. Conclusions

This paper discussed certain methodological issues around the constitutive rules account of assertion and by extension the institutional role of being a professor. The defenders of Williamsonian constitutive rules sometimes argue in favor of their view by pointing to the intuitive possibility of offering

the internal criticism of an assertion or of being a professor. I argued, first, that where the rules which the internal criticism appeals to are implicit, it is difficult to intuitively distinguish the internal criticism from other normative sources of justification, e.g., generic social or moral norms. Second, I showed how competing accounts, such as the attitudinal account of assertion or the regulative rules account, can account for the objections drawn from intuitions about the possibility of internal criticism. While the point of these arguments was not to directly establish the truth or falsehood of any single account, in order to do so it is necessary to be clear about the evidence that can be used for deciding these matters, which was the aim of the present paper.

Tampere University

References

Bach, K. and R. Harnish (1979). *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Epstein, B. (2015). *The Ant Trap*. Oxford: Oxford University Press.

Goldberg, S. (2015). *Assertion*. Oxford: Oxford University Press.

Grice, P. (1957). "Meaning." *Philosophical Review* 66, no.3, 377–388.

Guala, F. and F. Hindriks (2015). "A Unified Social Ontology." *Philosophical Quarterly* 65, 177–201.

MacFarlane, J. (2011). "What Is Assertion?" In J. Brown and H. Cappelen (eds.), *Assertion: New Philosophical Essays*. Ed. Oxford: Oxford University Press, 79–96.

García-Carpintero, M. (2021). "How to Understand Rule-Constituted Kinds." *Review of Philosophy and Psychology* 13, 7–27.

Roversi, C. (2021). "In Defence of Constitutive Rules." *Synthese* 199, 14349–14370.

Williamson, T. (1996). "Knowing and Asserting." *Philosophical Review* 105, 489–523.

What Could and What Should Be Said? On Semantic Correctness and Semantic Prescriptions

ALEKSI HONKASALO

1. Introduction

The thesis that meaning is normative is the claim that there is an essential normative component in meaning. This essential component has been linked with the intuitive classification of language use in terms of correct and incorrect applications; an English speaker does something correct when she uses the word "apple" to refer to apples and something incorrect when she uses the word "orange" to refer to apples.[1] While the claim that meaning is normative used to be taken as trivial, in the last two decades this thesis has garnered a significant amount of criticism. Anti-normativists—as they are sometimes referred to in the literature—claim that while the notion of semantic correctness is necessarily tied to meaning, normativity is not. While both sides of the debate agree on the existence of the semantic correctness conditions, defenders of the normativity thesis see this classification of actions as a normative feature of meaning while the opponents claim that semantic correctness is not normative, arguing that it does not tell the speaker what she ought to do.

The aim of this paper is to investigate the relationship between semantic correctness and the normativity of meaning understood here in terms of prescriptions, which tell speakers what they ought to do in certain circumstances. I will distin-

[1] I will use quotation marks to signify words and italics to refer to semantic content. E.g., "green" (word) means *green* (semantic content) and refers to green things (entities). The italics could be compared to David Kaplan's meaning marks. (D. Kaplan 1968, 186; Kripke 1982, 10, footnote 8).

guish two questions: (1) Can semantic correctness be accounted for without also providing an account of semantic prescriptions? (2) Can semantic prescriptions be derived from correctness? I will attempt to elucidate how these two questions relate to each other, by distinguishing two construals of the thesis: the metasemantic construal and the metametasemantic construal, the first of which is a thesis about meaning, second is a thesis about theories of meaning. A negative answer to question (1) corresponds to the metametasemantic construal; all theories of meaning must provide some kind of account for both semantic correctness and semantic prescriptions. An affirmative answer in turn implies the failure of the metametasemantic construal; at least some account of meaning can be provided without semantic prescriptions.

However, the failure of the metametasemantic construal of the thesis does not imply the failure of the metasemantic one. For semantic correctness to be prescriptive in the metasemantic sense, it suffices that at least some plausible theories of meaning can treat semantic correctness as prescriptive. In contrast, the negative answer to the second question is the anthisesis of the metametasemantic thesis; all plausible theories of meaning must reject semantic prescriptions. The crucial difference between these two readings is that the metametasemantic thesis must be decided without making assumptions about the nature of meaning that go beyond the pretheoretic concept of meaning. The metasemantic thesis, on the other hand, can more freely invoke more substantial assumptions about the nature of meaning.

Kathrin Glüer and Åsa Wikforss have argued that the answer to (1) is affirmative. They argue that since correctness can be understood non-prescriptively, semantic correctness does not entail semantic prescriptions (Glüer and Wikforss 2009; 2015). Similarly, Anandi Hattiangadi has argued that semantic prescriptions require speakers to speak the truth and thus cannot be semantic in nature. Only plausible "oughts" derivable from semantic correctness are dependent on the speaker's desire to speak the truth or to communicate and as such are merely hypothetical prescriptions, which fail to show that meaning is normative. (Hattiangadi 2006; 2007; 2009) If this is correct and no plausible prescriptions can be

derived from the semantic correctness conditions, the answer to question (2) must be negative.

Some of the normativists reject these arguments while others only disagree with some of the anti-normativists' claims. Daniel Whiting has argued that if correctness is taken as a higher-order feature, the anti-normativists arguments fail to show that semantic correctness can be understood non-prescriptively. He also argues that the problems of semantic prescriptions can be circumvented by reformulating prescriptions by using "may" rather than "ought" and by relying on the idea that an agent's obligations can be overridden by other obligations. (Whiting 2007; 2009; 2016.) Claudine Verheggen, on the other hand, agrees with Hattiangadi that semantic prescriptions are contingent on the speaker's desires but argues that they are still essential to meaning (Verheggen 2011). Finally, Alan Millar distinguishes two notions of semantic correctness and argues that while one of these is prescriptive, the other is not (Millar 2002; 2004; see also Buleandra 2008; Reiland 2023).

In this paper, I will argue against the normativity of meaning, both as a claim concerning meaning and as a claim concerning theories of meaning. I will first defend the claim that semantic correctness can be understood non-prescriptively. This shows that the normativity of meaning cannot act as a criterion of adequacy for plausible theories of meaning. However, the failure of the metametasemantic thesis does not settle the question of whether correctness can also be understood prescriptively. In the latter part of the paper, I will discuss the problems faced by possible formulations of semantic prescriptions and argue that these problems diminish the plausibility of normative theories of meaning that do interpret semantic correctness as prescriptive.

I will argue that the semantic prescriptions, advocated by Whiting, demand speakers to use expressions that are unsuitable for expressing what they want to express. The fact that these prescriptions ignore speakers' communicative intentions this way shows that they cannot be semantic in nature. Furthermore, a plausible candidate for semantic prescriptions would depend on what speaker's communicative intentions are, and therefore cannot be derived from the commonly accepted notion of semantic correctness alone.

While the considerations above might suggest that it is the non-prescriptivity of correctness that acts as a criterion of adequacy for theories of meaning, I will refrain from drawing a conclusion this strong, since the possibility of plausible candidates for semantic prescriptions, cannot be entirely ruled out on the basis of this paper alone. A further study of alternative notions of semantic correctness advocated by Millar and others would be required before the reversal of the metametasemantic claim can be accepted. However, even if a plausible set of prescriptions could be found, non-prescriptive theories might still be overall preferable. The appropriateness of expressions for communicative intentions might be better captured by identifying what the expression can be used for without taking a stance on what it may or may not be used for.

I will begin in section 2 by characterizing the notion of semantic correctness and discuss how it relates to the normativity of meaning and discussions concerning the naturalizability of meaning. In section 3, I will turn to the discussion on whether the general notion of correctness can be understood non-prescriptively. The next two sections concern which prescriptions could be the semantic prescriptions if semantic correctness is assumed to be prescriptive. In section 4, I will show that Whiting's formulations, which do follow from the assumption that semantic correctness is prescriptive, are in conflict with speakers' communicative intentions and sketch an anti-normativist account of what it is to act according to communicative intentions based on what can be done not what should be done. In the section 5, I will consider some normativist alternatives that aim to take into account what speakers want to express. In particular, I will focus on proposals by Claudine Verheggen (2011; 2015) and Alan Millar (2002; 2004). I will argue that these would be better understood in terms of non-semantic prescriptions as well.

2. Semantic correctness, normativity, and naturalism

Following Kripke's (1982) discussion on the rule-following, many philosophers were keen to adopt the slogan "meaning

is normative." It was generally agreed that normativity played a key role in Kripke's arguments against various theories of meaning and in particular the naturalized ones.[2] Furthermore, rather than being a feature ascribed to meaning by some theories of meaning, the slogan was taken to capture a pretheoretical criterion of adequacy for the theories of meaning. In other words, the normativity of meaning is a metametasemantic thesis about what kind of theories can be acceptable theories of meaning. Although this distinction is not always made explicit, this is how the claim is typically formulated in the debate (e.g., Glüer and Wikforss 2015, 64; Hattiangadi 2006, 220; Verheggen 2011, 553; Whiting 2009, 553).

The early advocates of the normativity thesis remained divided on how exactly the normativity should be understood. Paul Boghossian (1989) critically discussed many of these early reactions to Kripke as well as presented one of the most influential intepretations of the normativity of meaning in terms of semantic correctness conditions. The normativity of meaning, according to Boghossian, is just the uncontroversial claim that the world "green" applies correctly to green and only green things. While virtually nothing in philosophy is entirely free of controversy, Boghossian's understanding of normativity is of special interest, since today many anti-normativists are willing to accept the claim that meaningful expressions have correctness conditions (Hattiangadi 2006, 222; Glüer and Wikforss 2015, 66).

What anti-normativists deny, however, is that the semantic correctness conditions do anything beyond categorizing utterances into correct and incorrect ones. The existence of such categorizations is not sufficient to show that meaning is normative at least in the sense that threatens naturalized theories of meaning.[3] (Hattiangadi 2006, 222; see also Glüer and

---

[2] The role of normativity in Kripke's arguments has also been questioned; see Kusch 2006. Although I don't intend to endorse it, I will sometimes use a normativist reading of Kripke to elucidate the supposed intuitive link between normativity of meaning and correctness.

[3] Like normativity, naturalism is a notoriously ambiguous notion. To borrow Papineau's (2006) rough characterization, naturalized theories explain concepts like meaning without extending the methods and ontology of the natural sciences.

Wikforss 2015, 66.) Suppose we accept that Kripke's considerations show that some naturalized theories, namely dispositionalist theories of meaning, which identify meaning facts in terms of speakers' dispositions to use expressions in certain ways, fail to establish the correctness conditions of expressions. It is less clear, however, that the same arguments can be applied to more sophisticated versions of dispositionalism or theories relying on facts beyond dispositions such as speaker's causal history or biological functions to identify the meaning facts.

Anti-normativists argue that a further assumption is needed to show that semantic correctness also presents a problem for the more sophisticated theories. Semantic correctness must also be shown to be prescriptive. Only then it could be argued that naturalized theories of meaning illegitimately derive "ought" statements from "is" statements. (Hattiangadi 2006, 222–24; 2007, 35, 37; Glüer and Wikforss 2009, 32.) Intuitively dispositionalism merely describes how a speaker uses expressions and not how they should be used. Perhaps the situation is similar with the other naturalist theories of meaning. After all, describing what is the case is the aim of scientific inquiry, not prescribing what should be the case. Even if, say, the speaker's causal history with the concept "green" can offer a candidate classification of utterances into correct and incorrect ones, it will ultimately fail to explain the entailed semantic prescriptions.

It is worth stressing that anti-normativists do not argue that normativity in general is naturalizable. Nor is the goal to offer a naturalistic analysis of semantic normativity. Rather anti-normativists argue that there is no semantic normativity to be naturalized beyond perhaps the trivial correctness which any theory can account for. If no prescriptions follow from semantic correctness, there are no normative truths for a theory of meaning to explain. Therefore, normativity does not justify extending the scope of Kripke's argument beyond simple dispositionalism regardless of whether normativity can be naturalized or not.

Not everyone agrees with this evaluation, however. Claudine Verheggen claims that the core problem for the semantic naturalist is not explaining the prescriptions implied by semantic correctness. Rather the core problem of naturalization,

according to her, is to explain semantic correctness itself (Verheggen 2011, 556). Likewise, Jeffrey Kaplan has argued that even if semantic correctness was not prescriptive, this would not mean that it has to be descriptive (J. Kaplan 2020).[4] Additionally, correctness might have some other normative implications beyond prescriptions. If there are some non-prescriptive, but still normative implications, showing that no prescriptions follow from correctness may not be enough to show that semantic correctness is not normative; only that semantic correctness is not prescriptive.

Going through all possible normative implications of semantic correctness would be beyond the scope of this paper. I will therefore limit the study to prescriptions and accordingly shift the terminology from normativity to prescriptivity. Given that ought is a central normative term, failing to imply semantic prescriptions could still reflect deeper issues with the normativity of meaning. Nevertheless, the categorical conclusion that meaning is not normative cannot be drawn based on this paper alone.

3. Is correctness necessarily prescriptive?

Before discussing the notion of semantic correctness, it is worthwhile to examine the relationship between prescriptivity and the general notion of correctness. Whiting and Jaroslav Peregrin take "correct" to be a part of the basic normative vocabulary among "ought," "may," "obligation," and "permission" (Whiting 2009, 538; Peregrin 2012, 84). They argue that since correctness is an intrinsically normative notion, prescriptions do follow from correctness. Why then anti-normativists reject this intuition?

Glüer and Wikforss give two reasons to think that semantic correctness is not necessarily a prescriptive notion. First, they suggest that given that semantic correctness is a technical philosophical concept, the facts about natural language usage of the word "correct" offer only limited philosophical

[4] If a theory has problems accounting for correctness itself these may simply be symptoms of a more substantial issue with the theory. That is, the theory fails to account for correctness because it fails to give a plausible account of meaning and not the other way around. (See Honkasalo 2022.)

import. Instead, "correctness" should be understood as a placeholder term to be replaced by the basic concept of the semantic theory, such as truth, which need not be normative. Second, they contend that the word "correct" does have some non-normative uses, namely, conforming to a standard, which need not entail prescriptions. They conclude that, unless there is an additional argument to support the normativist claim, semantic correctness can be understood merely as a categorization of applications into correct and incorrect without prescriptive implications. While applying the word "green" to a red entity does not conform to the correctness conditions of the word "green," this does not straightforwardly imply that the applications should be corrected or frowned upon. (Glüer and Wikforss 2015, 68; see also Hattiangadi 2006, 222.)

However, Whiting claims that by treating the correctness as a placeholder, anti-normativists fail to recognize the distinction between the concept of correctness and the correctness-making feature. Relying on a distinction highlighted by Gideon Rosen, he argues that while the fact that the object to which the expression "green" is applied is a purely descriptive fact, this fact is merely the correctness-making feature that must obtain for the application to be considered correct. Claiming that an application of "green" is correct, on the other hand, is a higher-order claim that the application possesses the features required for it to be correct. (Whiting 2009, 538–39; Rosen 2001, 619–29.)[5]

However, pointing out the distinction only serves to move the question of normativity of correctness to a higher level, the fact of which Rosen is keenly aware. In order to argue that correctness is prescriptive, it is not enough to show that notions of correctness and correct-making feature are distinct (Rosen 2001, 620–21).[6] While Rosen is sympathetic to the idea

[5] Glüer and Wikforss claim that this would merely make it possible for the normativist to accept that the basic semantic concept is non-prescriptive, but maintain that correctness could still be prescriptive. This would not however be enough to show that correctness must be understood prescriptively in the higher-order sense. (Glüer and Wikforss 2015, 71)

[6] In (2001) Rosen is concerned with the normativity of belief rather than meaning. Regarding the relationship between correct and true belief, he writes: "it is not enough […] that correctness and truth should be distinct.

explicitly endorsed by Whiting and Peregrin that "correctness" in some sense could be counted amongst the normative vocabulary, according to him, correctness differs in one crucial way from typically normative terms like "ought," namely it lacks the "internal connection" with reasons for action.[7] According to Rosen, it is not enough to recognize that it is correct to play the note B in the second bar of the Piano sonata to motivate a (rational) person to play it (Rosen 2001, 620–21). Perhaps the player wishes to amuse the audience by intentionally playing the piece incorrectly or maybe she does not wish to play Mozart in the first place. The fact that a correct rendition of Beethoven's Moonlight Sonata is an incorrect recital of Mozart's Sonata does not mean that playing Beethoven should be avoided.[8]

What does this disconnect with reasons mean for the prescriptivity of correctness? Using the terminology favoured in the debate so far, this means that prescriptions implied by the correctness conditions are at most merely hypothetical prescriptions, which might also be called "technical norms," or "means-to-end prescriptions," that tell what an agent ought to do to achieve a goal. In contrast, categorical prescriptions tell an agent what to do regardless of what goals an

---

It remains to show that correctness is normative feature." It is however clear that the point is applicable to the case of meaning as well.

[7] Rosen (2001, 621), however, points out that in another sense correctness has more in common with normative vocabulary. Namely, in a sense that while we cannot say that one ought to play a musical piece correctly, we can in principle say from any recital whether the piece was played correctly or not regardless of what goals or desires a player may have. However, since the aim of this paper is not to show that any conception of the normativity of meaning is untenable, only that semantic prescriptions cannot be derived from semantic correctness, I will leave this problem aside.

[8] One might question whether the correctness conditions of Mozart's Sonata should be applicable to a rendition of Beethoven's sonata. However, if the further notion of applicability is needed, then the notion of correctness is not, in itself, sufficient to provide reasons for action. Furthermore, what source for the appropriateness there is other than players desire to play the piece or some extramusical obligation (such as a promise) to play it? For discussion on the notion of applicability in the context of the normativity of meaning debate, see Reinikainen 2020.

agent takes to be worthy of accomplishing. If semantic correctness necessarily entails such categorical prescriptions, theories of meaning that fail to account for them would indeed provide an incomplete picture of meaning, but if correctness only implies prescriptions that are contingent on speaker's goals or desires such a conclusion would be too hastily drawn.

To begin with, hypothetical prescriptions might only look like prescriptions, but instead, be equivalent to descriptive claims. The mere apparence of the word "ought" is not enough to guarantee that these are really prescriptions, since the word also appears in descriptive statements like "it ought to rain soon," which predicts rather than prescribes. Likewise, hypothetical prescriptions have been suggested to be merely descriptive claims in prescriptive disguise. According to R.M. Hare, the statement "If you want to go to the largest grocer in Oxford, [you ought to[9]] go to Grimbly Hughes" says nothing more than the statement: "Grimbly Hughes is the largest grocer in Oxford" (Hare 1952, chap. 3; see also Hattiangadi 2006, 228). More generally, hypothetical prescriptions could be interpreted as directions or recipes which describe which actions are sufficient for achieving a certain goal, instead of prescribing that those actions ought to be taken or saying anything about whether the goal is worth achieving.

However, it would also be too hasty to conclude that hypothetical prescriptions are necessarily just rephrased descriptive claims. Although he shared Hare's reservations about calling them prescriptive, von Wright was hesitant to identify hypothetical prescriptions with descriptive statements, since the descriptive claim about the largest grocer says nothing about anyone's mental states (von Wright 1963, 9–10). It is also important to note that even Hare treats *want* as a "logical term" in his analysis rather than as an ordinary term relating to mental states such as desires. According to him, if we instead interpreted the *want* to signify a mental state, the hypothetical prescription does say more than the descriptive

[9] Hare discusses imperatives rather than prescriptions and hence the original says only "go to." I have changed the imperative to an "ought"-statement to better suit the argumentation of this paper. The addition of "ought" in this case does not distort the intent of the original passage.

claim—namely that if you have a desire, or you have adopted a goal, you really ought to go act on your desire or a goal (Hare 1952). In this case, the *ought* is no different from categorical prescriptions; it is merely conditional on mental facts.

However, interpreting means-to-end prescriptions as conditional prescriptions would also produce a problem: we would ought to act on any desire or a goal and undertake the means, however immoral, in pursuing them (Hattiangadi 2006, 228). If I want to get an inheritance no matter the cost, should I serve arsenic at dinner in order to kill a rich relative? Moral conflicts aside, it has also been questioned whether anything truly normative could really be conditional on the adoption of a goal or a desire since this seems to make *oughts* too easy to come by (e.g., Bratman 1981; Broome 2013).

Fortunately for the purposes of this paper, we can leave these difficult questions open as well as leave various important issues unaddressed,[10] since regardless of the way we account for the means-to-end prescriptions, the normative status of correctness is left unaffected. If these prescriptions are interpreted as descriptions in disguise, they obviously provide no reason to think correctness is prescriptive. If, as von Wright suggests, they are not descriptive, but not prescriptive either, then we arrive at the same conclusion. Even if there is something genuinly prescriptive about means-to-end prescriptions, nevertheless correctness only determines the means and not the oughts of the prescriptions. They merely identify the notes which satisfy the goal of playing Mozart correctly, but something else prescribes that those notes ought to be played. This is because whatever prescribes a player to undertake the musical means to the musical ends must be what makes any means-to-end prescriptions prescriptive, most of which have nothing to do with Mozart or music in general.

What this means is that we can accept that if something is semantically correct, there is always a corresponding hypothetical prescription. We can even accept that these prescriptions are somehow genuinely normative, but still maintain

[10] Including issues such as: Do means-to-end prescriptions require actions to be performed or merely intended? Should I intend what I believe to be the means or which actually are the means?

that semantic correctness is not the source of normativity. In other words, if a speaker intends to use expressions correctly, she may be required to apply "green" only to green things but meaning requires no such thing. Assuming non-naturalism about normativity, the hypothetical *oughts* might themselves pose a problem for the naturalist philosopher, but this does not affect the naturalization of meaning. The only thing about these prescriptions a theory of meaning needs to explain is how to behave semantically correctly and not that one should behave so. The latter would be like requiring toxicology to explain not only that arsenic is poisonous, but also why one ought not to feed it to one's guests.

Based on the considerations presented in this section, it seems that the notion of semantic correctness itself does not provide a straightforward argument for the metametasemantic thesis. The normativity of meaning cannot be a criterion of adequacy for plausible theories of meaning based on semantic correctness alone, since correctness can be understood non-prescriptively. An advocate of a naturalized theory of meaning can accept that there are correctness conditions but deny that they generate any special semantic prescriptions to be accounted for.

Of course, the fact that meaning can be understood non-normatively does not imply that it cannot be understood normatively. Neither does the failure of the metametasemantic thesis settle the question of whether meaning is actually normative, since the thesis concerns only which theories manage to capture pretheoretical constraints, not which theory is true. If meaning is actually normative, the fact that a naturalist reductionist theory of meaning captures pretheoretical intuitions is an uninteresting consolation prize. Additionally, out of all plausible theories of meaning, it might be the case that the best theories of meaning do imply that semantic correctness is prescriptive.

In the next section I will shift the attention to the question of whether semantic correctness can be understood prescriptively. To assess this matter we must take a closer look at what prescriptions could be said to follow from semantic correctness. I will argue that semantic correctness cannot plausibly determine what a speaker ought to do purely from the point of view of meaning. While this may not be sufficient for

establishing the reversal of the metametasemantic thesis—i.e., it may not be sufficient to show that all plausible theories of meaning must interpret semantic correctness non-prescriptively—it nonetheless suggests that the non-prescriptive alternative is preferable since it avoids these issues simply by leaving the question of what a speaker should do with words to be determined by something other than meaning.

4. Should you speak correctly?

If we suppose that semantic correctness does entail semantic prescriptions, how should these correctness conditions and prescriptions be formulated? Hattiangadi (2006) formulates the correctness conditions of application in the following manner. Let t be a term, F be the meaning associated with t, and f be a feature or collection of features that make it the case that F applies:

(CA) t means $F \rightarrow (\forall x)(t$ applies correctly to $x \leftrightarrow x$ is $f)$.

The expression "green" means *green* which applies to entities that are green, therefore the expression "green" applies correctly to green entities and incorrectly to non-green ones. A straightforward way to capture the intuition that semantic correctness is prescriptive is to require speakers to use expressions correctly. To represent this we can modify our prescription schema by replacing the phrase "S applies correctly" with "S ought to apply," where S is a speaker.

(SP1) t means $F \rightarrow (\forall x)(S$ ought to apply t to $x \leftrightarrow x$ is $f)$.

As Hattiangadi points out, (SP1) requires too much from the speaker. Suppose that there is a dog on Mars. It follows then that a speaker ought to call it a dog regardless of whether she is aware of its existence. (Hattiangadi 2006, 226–27.) Moreover, since (SP1) is formulated schematically, a speaker ought to apply a proper name to its bearer and state every property it instantiates. (SP1) then clearly violates the principle of *ought implies can*.

However, (SP1) is not the only possible option to capture the prescriptivity of correctness. Peregrin and Whiting suggest that switching "ought" to "may" better captures the idea (Whiting 2009, 544–45; Peregrin 2012, 87–88). After all, mov-

ing a bishop diagonally is a correct move, but this does not imply that this move ought to be made, since one can also make another correct move. Similarly, perhaps the semantic prescriptions are best captured by the schema:

(SP2) t means $F \rightarrow (\forall x)(S \text{ may apply } t \text{ to } x \leftrightarrow x \text{ is } f)$.

(SP2) is no longer in a straightforward conflict with the principle of *ought implies can*. If there is a dog on Mars, a speaker may apply "dog" to it, but it no longer follows that she ought to do so. One might be concerned whether *may* is strong enough to constrain speakers' actions for us to consider (SP2) as a genuine prescription? This, however, is not an issue, since in addition to telling speakers what may be done (SP2) also implies what the speaker ought not to do, namely, she ought to refrain from applying "dog" to non-dogs. (SP2) is, therefore, more accurately called prohibition rather than permission.

While (SP2) no longer contradicts the *ought implies can* principle, Hattiangadi points out it may nonetheless contradict other obligations a speaker may have. Sometimes a speaker may be morally obligated to lie and therefore speaker ought to apply t to x and she also ought not to apply t to x.

Whiting does not see such a contradiction as a serious problem. He accepts that some other normative obligations (moral, epistemic, or prudential) can be in conflict with speakers' semantic obligations. For meaning to be normative it suffices that meaning provides a reason for not applying t to x, even if there are weightier reasons for applying t to x. In other words, (SP2) is a *prima facie* prescription or a prescription that can be overridden by other obligations. Another way of putting this is to say that the fact that t means F is a *pro tanto* reason for applying t to only things that are f's even if all reasons considered one ought to apply t to an entity that is not f. To characterize Whiting's view of normativity more informally: if there are no weightier reasons to do otherwise, a speaker ought to refrain from applying "dog" to non-dogs. (Whiting 2009, 546.)[11]

[11] Peregrin raises a similar point by distinguishing defeasible/indefeasible obligations. However, the core issue is the same: how to explain conflicting obligations (Peregrin 2012, 80). Therefore, I take it that Peregrin's ob-

In an anticipation of this kind of defense, Hattiangadi claims that while *prima facie* obligations can only be overridden by other obligations, semantic prescriptions like (SP2) seem to be overridable by mere desires (Hattiangadi 2006, 232). Namely, if a speaker has no interest in telling the truth and instead wishes to tell a lie or a fictional story, there does not seem to be a semantic reason to criticize her linguistic behaviour. Therefore (SP2) cannot be regarded even as a *prima facie* prescription, since if all that is needed for excusing speaker's apparent transgressions is the fact that she just did not feel like abiding by it, then there was no transgression to begin with. The only other option is to maintain that the prescription is a hypothetical prescription and contingent on the desire to speak the truth, which—as we saw in the previous section—is not sufficient for Whiting's goals.

Whiting still maintains that even if a speaker had no desire to speak the truth, her behaviour may still be criticizable from a semantic perspective. A speaker does have a semantic reason not to apply dog to non-dogs and that reason does not cease to be a reason even if she has no desire to tell the truth. (Whiting 2009, 548–49) He, however, stresses that the fact that the speaker's behaviour is criticizable does not mean that her transgressions are particularly grievous. Semantic offenses are not on par with moral or epistemic offenses. He suspects that, at least partly, the source of anti-normativist apprehension towards the thesis is in taking the thesis to be stronger than it needs to be. Recognizing "the bearable lightness of meaning" can bring the thesis into a more favourable light. (Whiting 2009, 550–51; see also 2007, 139)

Hattiangadi claims however that even if there are no reasons to act otherwise, (SP2) is still contingent on the desire to communicate. If a speaker has no intention to communicate at all, what reason is there to criticize her behaviour? (Hattiangadi 2006, 232) However, I do not think that this is the core issue with (SP2). First, if a speaker lacks the desire to communicate, then we could reasonably question whether the speaker simply does not speak English or any other lan-

---

jection can be formulated in terms of *prima facie* obligations or *pro tanto* reasons as well.

guage.[12] Since (SP2) does not generate obligations for those who are not using language, if the speaker ceses to speak English, it would indicate that Hattiangadi's proposed counterexample does not have a bearing on the plausibility of (SP2) which must be assessed by keeping the meaning condition fixed.

Secondly, if Jane calls a cat "dog" just because she felt like lying, she must have had a desire to communicate. Namely, she has a desire to communicate a false proposition that a cat is a dog. Therefore the lack of desire to communicate does not affect whether or not (SP2) is in force for her. However, recognizing her intentions to communicate seems to indicate that despite her use being semantically incorrect, Jane did something right since she used precisely the right word given her communicative intentions and, therefore, in an accordance with the meaning of the word "dog."[13] While Whiting claims that since her application was semantically incorrect, she must have done something semantically criticizable if her choice of words corresponds with what she wanted to say, why should we take her application to be criticizable on semantic grounds? In contrast, if she applies "cat" to a cat, then she used a word that did not suit her intentions. If anything is criticizable on semantic grounds here, then should it not be this application even if it was the correct one?

One might try to argue that speakers' communicative intentions when lying should be regarded as a special case be-

---

[12] This response assumes that one might simply by forsaking any desires to communicate cease to speak English while making sounds that bear striking resemblance to English words. It is of course a non-trivial assumption that one might simply decide to opt-out of speaking a public language. Nevertheless, since the question whether or not the speaker speaks English can only affect the meaning condition in (SP2), it does not have a bearing on the question of what should be done in the case "dog" means *dog* for the speaker and therefore the choice does not have direct impact on the plausibility of (SP2). For discussion on public language, see Reiland 2021.

[13] Perhaps her use can be characterized even as semantically correct in Millar's sense (2004). For now, however, I will focus on Whiting's preferred notion of correctness as captured by (CA) and return to Millar's formulation in the next section.

cause lying is somehow parasitic on truthtelling,[14] but (SP2), taken as a semantic prescription, is not even contingent on speaker's desire to speak the truth. Suppose John wants to tell the truth, but mistakes a cat for a dog and therefore calls it "dog." While refraining from applying "dog" to a non-dog would be required to fulfill his desire to speak the truth, the proposition he intended to express was not the truth, but rather what he believed to be the truth. So it seems that John had two intentions, to speak the truth and to call an entity a dog. While his choice of words failed to satisfy the first intention, they reflected the latter intention adequately. Only the latter intention is relevant to assess whether he used "dog" in accordance with its meaning and therefore it should also be relevant when assessing if his behaviour warrants criticism from purely semantic perspective.[15]

Even though they might have done what their communicative intentions require, Whiting denies that Jane and John did something they *semantically* ought to have done. Instead, what makes their word choices successful is just the appropriateness for their communicative intentions. Any requirements Jane and John may have fulfilled are therefore contingent on their intentions and should be accounted for in terms of means-to-end prescriptions which, as seen in section 3, do not generate *oughts* of semantic kind (Whiting 2016, 229). While I agree with Whiting's assessment, the question remains whether he can also maintain that Jane and John also did something that from the *semantic* perspective they should not have done, since they failed to use words in an accordance with the correctness conditions.

He invites us to consider an analogy to chess where mistaking a rook for a bishop might explain why a player moved a piece diagonally, but even though the epistemic mistake might explain player's actions it does not change the fact that the move was against the rules of chess. Whiting maintains that similarly, John's failure to recognize a cat as a non-dog does not change the fact that (I) what he did was *semantically* incorrect and therefore (II) what he *semantically* ought not to

---

[14] Hattiangadi also considers this option, but rejects it (2006, 230–31).

[15] Wikforss (2001, 205–6) argues similarly that semantic prescriptions are ill-equipped to deal with reporting false beliefs.

have done (Whiting 2016, 232). Furthermore, Whiting stresses that (III) even if our account of meaning fails to recognize semantic mistakes as semantically forbidden, that does not mean this notion escapes the analysis. Just as in chess, where we can distinguish violations of the rule which are explained by player's mistake about the rule or which piece is which and cases in which a player intentionally breaks the rules, we can distinguish epistemic mistakes about the species of an observed animal from semantic mistakes about the meaning of the word "dog." Recognizing these type of semantic mistakes does not warrant the acceptance of additional semantic prescriptions, which forbid semantic mistakes. (Whiting 2016, 233–34.)

I agree with Whiting on points (I) and (III), but I am still inclined to deny (II). In the case of chess players' mental states like desires, beliefs, or intentions do not factor into deciding which moves are correct and incorrect or how pieces may or may not be moved.[16] While the same can be said about semantic correctness, the same cannot be said about the alleged semantic prescriptions. If there are semantic prescriptions at all, then what proposition the speaker wants to express should have a bearing on what she semantically ought to do. While the phenomenon of semantic mistake can be accounted for without invoking semantic prescriptions, the prescriptions which treat actions that are not semantic mistakes as semantically forbidden should also be regarded as non-semantic.

Whiting might contest this intuition and maintain that the prohibitions against incorrect speech are essential to meaning whereas semantic mistakes, despite their name, are at heart still factual mistakes, that is, mistakes about the true meaning of a word (Whiting 2016, 233–34). However, if we understand semantic mistakes as failing to use an appropriate expression for what speaker wants to express, this does not itself depend on speaker's beliefs. Mary might know the meaning of a

[16] One might object that whether or not you ought to follow rules of chess is contingent on the desire to play chess. I will not discuss this issue here, since if rules of chess generate merely hypothetical prescriptions, the analogy would support the anti-normativist rather than the normativist conclusion.

word "dog" and recognize the animal, but by the slip of a tongue call it "log." The appropriateness analysis of a semantic mistake therefore does not necessarily involve a factual mistake and this is what counts in favour of adopting it. However, even if the analysis fails to capture the distinction the examples of Jane, John, and Mary are nevertheless categorized respecting pretheorethical intuitions of semantic mistakes which is itself sufficient to favour theories that reject semantic prescriptions which altogether ignore speakers' intentions. Perhaps we could go as far as to claim that these theories should fail to be adequate theories of meaning themselves.

5. What is it that you want to say?

The moral of the last section was that (SP2) ignores what speakers want to express by their utterances and by doing so it permits some actions intuitively characterized as mistakes such as mistakenly telling the truth when attempting to lie, and forbids some actions which do not seem to call for semantic criticism, such as reporting false beliefs. Even if the last section is enough to justify rejection of some candidate prescriptions, the question remains whether some other prescriptions could fare better?

A worry might arise that in invoking the notion of semantic mistake, we ended up introducing another normative term that might imply semantic prescriptions. However, in the last section the mistakes were identified in terms of a mismatch between what the speaker wants to express and with which expressions she attempts to achieve these goals. What is left for the theory of meaning is to explain which expressions are suitable for the speaker's communicative intentions. Explaining what expression a speaker ought to use is and indeed should be regarded as something beyond its scope. To put this concisely, a theory of meaning must explain how expressions can be used, not how they ought to be used.

Crucially, a prescription candidate which would better capture our intuitions on semantic mistakes cannot be derived from semantic correctness, as it is formulated in (CA). After all, the categorization of correct and incorrect applications does not coincide with cases that can intuitively be

characterized as mistakes. Therefore, a more adequate prescription would need to deem some correct uses as ones to be avoided and some incorrect uses to be accepted. No simple argument is available which shows that (CA) entails such alternative prescriptions. Indeed, it is not easy to see what argument could show that semantic correctness is prescriptive notion, but sometimes you ought to behave correctly and other times incorrectly. Since this is essentially what it takes to capture the semantic intuitions, then no alternative schemas can fair any better.

Nevertheless, before calling semantic correctness non-prescriptive I need to address some counterpoints. Normativists have at least two ways of countering the reasoning above. First, it could be argued that the criteria of prescriptivity should be relaxed. Although Whiting is ready to accept the anti-normativist claim that categorical prescriptions are what is needed for meaning to be genuinely normative, some, such as Verheggen, are not so quick to dismiss hypothetical prescriptions. Secondly, it has been argued that (CA) does not capture the intended semantic correctness and that the right formulation of semantic correctness could imply categorical prescriptions (Buleandra 2008; Millar 2004; Reiland 2023).

Verheggen accepts that no categorical prescriptions can be derived from correctness conditions since whether a speaker ought to apply a word to an entity or not depends on how she wishes to employ it. However, correctness conditions prescribe how to employ the words when you want to be sincere, nonsincere, or humorous. She, however, claims that while this makes the prescriptions hypothetical, they are not analogous to means-to-end prescriptions, which can arise from any fact, because these prescriptions are essential to meaning. She points out that facts about rain and umbrellas are just the same whether I want to stay dry or not. If I do not mind getting wet, then these facts simply become irrelevant for considering what to do. She argues that, while prescriptions implied by semantic correctness are dependent on the speaker's desires, they do not become irrelevant even if those desires change. This is because regardless of what the speaker wants to say, correctness conditions imply what they ought to do in that circumstance. If the correctness conditions of the

word "dog" become irrelevant for the speaker, it can only be because she does not mean anything by "dog." (Verheggen 2011, 562–63)

I agree that the hypothetical prescriptions are entangled with meaning facts, but I disagree that this would show that there is a disanalogy with the means-to-end prescriptions. All this entanglement amounts to is that meaning determines the means regardless of what ends speakers have, and this does not make the "oughts" essential to meaning. According to the picture I have advocated here, the reason for this entanglement is that meaning of an expression determines what can be expressed with it and this is naturally tied with what actions are required for attaining the speakers' intentions. In other words, even if the semantic correctness conditions of the word are necessary to determine the means-to-end prescriptions associated with that word, this does not mean that semantic correctness is also sufficient to entail those means ought to be undertaken. Something else must be the source of the "oughts" and the source must be common to all means-to-end prescriptions regardless of whether or not they have anything to do with meaning.

Moving on to the worry concerning the proper formulation of semantic correctness. Note that the problems of semantic prescriptions discussed so far stem from the close relationship between (CA) and truth. Therefore, it is no wonder that prescriptions do not condone false uses. Alan Millar recognizes this and distinguishes the correctness of application which corresponds to (CA) from the correctness of use. According to him, while the correctness conditions of applications do not in themselves prescribe actions in the sense that speakers's would be required or allowed only to utter correctly, they determine the conditions of correct use or use in an accordance with meaning.[17] (Millar 2004, 166–67.)

---

[17] Similarly, Reiland distinguishes referential correctness (which corresponds to the correctness of application) from linguistic correctness which is use in accordance with meaning. However, he claims that the notion of use in an accordance with meaning admits to both normativist and anti-normativist construal (Reiland 2023, 2198, fn. 7). Ruling out the possibility of normativist construal Reiland has in mind would require a more detailed treatment of his views. Such treatment is better offered in the context of a different paper.

In which conditions the speakers' uses accord with meaning? Without overly simplifying Millar's account we can characterize it in terms of absence of semantic mistakes. However, if mistakes are identified in terms of mismatch between what the speaker is intending to say and what expressions she uses, respecting correctness conditions of application just ends up implying hypothetical prescriptions and therefore the alternative notion of correctness offers no improvement compared to (CA).[18] Indeed Millar recognizes an alternative picture where prescriptions to use "dog" in certain ways might be contingent on speakers' intentions. However, he argues that this alternative would still have to assume that there is a background practice of meaning *dog* by "dog." Since practices are inherently normative for Millar, the prescriptions relating to correct use are in fact intrinsic to meaning, because the source of those prescriptions is in the practice of meaning. (Millar 2004, 167, 172.) However, the practice of meaning *dog* by "dog" itself could be accounted for in terms of "dog" being used to mean *dog*. This analysis, on the face of it, requires no additional prescriptions beyond the hypothetical ones.

Nevertheless, even if prescriptions are not required by the analysis does not mean they cannot be in force. That is, in addition to the means-to-end prescription there might also be a semantic prescription with identical requirements. These semantic prescriptions may not go against the pretheoretical intuition on the nature of meaning and therefore, pending a more detailed analysis of Millar's account, we cannot conclude the reversal of the metametasemantic claim—that meaning is not and cannot be normative—should be adopted. However, while there is nothing inherently wrong with having normative redundancy, these semantic prescriptions appear to offer no further insight into meaning, because their content is already captured by the means-to-end prescriptions. Theories that reject these prescriptions (ceteris paribus) would be simpler and therefore at least in some sense preferable.

---

[18] Whiting (2016, 229–30) also argues that correctness of use implies merely hypothetical prescriptions.

6. Concluding remarks: What is wrong with the normativity of meaning?

In this paper, the claim that semantic correctness is prescriptive was given two readings: metametasemantic and metasemantic. According to the metametasemantic reading, all plausible theories of meaning must interpret semantic correctness prescriptively. According to the metasemantic reading, semantic correctness is prescriptive, but this does not imply that all theories of meaning denying this automatically failed to capture the pretheoretical concept of meaning. A defense of the metasemantic claim can therefore depend on some substantial assumptions about meaning that go beyond the pretheoretical notion. In section 3, I defended the anti-normativist claim that since the general notion of correctness is not automatically prescriptive, the semantic correctness can be understood non-prescriptively. Because semantic correctness can also be interpreted non-prescriptively, anti-normativists are free to reject any prospective semantic prescriptions while maintaining that semantic correctness itself is essential to meaning.

In sections 4 and 5, I argued that semantic correctness in its simplest form (CA) cannot be prescriptive. If it were, some uses which intuitively warrant no semantic criticism would nonetheless be semantically forbidden. The only way to maintain that semantic correctness is essentially prescriptive is to argue that some alternative notion of correctness is prescriptive. Even if this alternative notion of semantic correctness produces a plausible theory of meaning, which presupposes semantic prescriptions, this would not mean that the theory should be preferred over the ones which require no semantic prescriptions.

The problem with deriving semantic prescriptions from semantic correctness is not, as Glüer, Hattiangadi, and Wikforss have argued, that the implied prescriptions are contingent on the desire to tell the truth or desire to communicate. The crux of the problem is that meaning seems to only determine what can be expressed by an expression whereas alleged semantic prescriptions concern what should be expressed. If a theory of meaning manages to explain the relationship between words and world, it can explain what

speakers can do with meaningful expressions. If the word "dog" means *dog*, then it can be used in expressing propositions like *a dog wears a hat*, but it is entirely another question whether this proposition should be expressed.

Traditionally, it has not been the task of a theory of meaning to explain what people should do with words, and it is unclear why such a thing would be a good idea. Semantics provides a toolbox of meaningful expressions for speakers to use, not a script to be followed. Formulating prescriptions as prohibitions would only produce a script with a little room for improvisation, but it would still be a script nonetheless. It would be perhaps too far to suggest that the normativist had mistaken what *can be done* to what *may be done*, but perhaps they have failed to appreciate how many of the intuitions relating to the latter can equally well be captured by the former.

Tampere University

References

Boghossian, P. (1989). "The Rule-Following Considerations." *Mind* 98(392), 507–49.

Bratman, M. (1981). "Intention and Means-End Reasoning." *The Philosophical Review* 90(2), 252.

Broome, J. (2013. *Rationality through Reasoning*. The Blackwell/Brown Lectures in Philosophy 4. Chichester, West Sussex ; Malden, MA: Wiley Blackwell.

Buleandra, A. (2008). "Normativity and Correctness: A Reply to Hattiangadi." *Acta Analytica* 23(2), 177–86.

Glüer, K., and Å. Wikforss (2009). "Against Content Normativity." *Mind* 118(469), 31–70.

Glüer, K., and Å. Wikforss (2015). "Meaning Normativism: Against the Simple Argument." *Organon F* 22 (Supplementary Issue), 63–73.

Glüer, K., and Å. Wikforss (2020). "The Normativity of Meaning and Content." In Edward N. Zalta (ed.), *The Stanfod Encyclopedia of Philosophy*. Fall 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/meaning-normativity/.

Hare, R. (1952). *The Language of Morals*. Oxford: Clarendon Press.

Hattiangadi, A. (2006). "Is Meaning Normative?" *Mind and Language* 21(2), 220–40.

Hattiangadi, A. (2007). *Oughts and Thoughts Rule-Following and the Normativity of Content*. Oxford: Clarendon Press.

Hattiangadi, A. (2009). "Some More Thoughts on Semantic Oughts: A Reply to Daniel Whiting." *Analysis* 69(1), 54–63.

Honkasalo, A. (2022). "Is semantic correctness descriptive?" *Theoria* 88(5), 899–907.

Kaplan, D. (1968). "Quantifying In." *Synthese* 19(1/2), 178–214.

Kaplan, J. (2020). "The Problem with Descriptive Correctness." *Ratio* 33(2), 79–86.

Kripke, S. (1982). *Wittgenstein On Rules and Private Language*. Reprinted in 2004. Cambridge MA: Harvard University Press.

Kusch, M. (2006). *A Sceptical Guide to Meaning and Rules: Defending Kripke's Wittgenstein*. Chesham: Acumen.

Millar, A. (2002). "The Normativity of Meaning." *Royal Institute of Philosophy Supplement* 51 (March): 57–73.

Millar, A. (2004). *Understanding People: Normativity and Rationalizing Explanation*. Oxford : New York: Clarendon Press ; Oxford University Press.

Papineau, D. (2006). "Naturalist Theories of Meaning." In Ernie Lepore and Barry C. Smith (eds.), *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press, 175–88.

Peregrin, J. (2012). "Inferentialism and the Normativity of Meaning." *Philosophia* 40(1), 75–97.

Reiland, I. (2023). "Linguistic Mistakes." *Erkenntnis* 88(5), 2191–2206.

Reinikainen, J. (2020). "Meaning Still Not Normative: On Assessment and Guidance." *International Journal of Philosophical Studies* 28(4), 510–26.

Rosen, G. (2001). "Brandom on Modality, Normativity and Intentionality." *Philosophy and Phenomenological Research* 63(3), 611–23.

Verheggen, C. (2011). "Semantic Normativity and Naturalism." *Logique et Analyse* 54(216), 553–67.

Verheggen, C. (2015). "Towards a New Kind of Semantic Normativity." *International Journal of Philosophical Studies* 23(3), 410–24.

Whiting, D. (2007). "The Normativity of Meaning Defended." *Analysis* 67(2), 133–40.

Whiting, D. (2009). "Is Meaning Fraught with Ought?" *Pacific Philosophical Quarterly* 90(4), 535–55.

Whiting, D. (2016). "What Is the Normativity of Meaning?" *Inquiry* 59(3), 219–38.

Wikforss, Å. (2001). "Semantic Normativity." *Philosophical Studies* 102(2), 203–26.

von Wright, G. H. (1963). *Norm and Action: A Logical Enquiry*. International Library of Philosophy and Scientific Method. London: Routledge & Kegan Paul [u.a.].

Rearticulated Psychological View of Generics and Worldly Truthmakers

PASI VALTONEN

1. Introduction

Generics are sentences like

(1) Tigers are striped.

(2) Ducks lay eggs.

(3) Mosquitoes carry the West Nile virus.

The curious thing about generics is that we generally think that they are true even though we know that there are counterexamples to them. We know that there are stripeless tigers. Only about half of ducks lay eggs, and less than 1% of mosquitoes actually carry the virus. Yet we think that (1), (2), and (3) are true. In contrast,

(4) Ducks are female.

(5) Bees are sterile.

are false. Surprisingly, the very same set that makes (2) true, fails to make (4) true. Concerning (5), over 90% of bees actually are sterile. Thus, (4) and (5) make it even more puzzling why we think that the set from (1) to (3) is true.

The accounts of generics are predominantly semantic. These treatments aim to come up with truth conditions for generics in a systematic way. The semantic accounts aim to explain compositionally why (2) is true and (4) is false. In contrast, Sarah-Jane Leslie argues that a psychological view can explain generics better. She thinks that generics do not have compositional truth conditions at all. Rather, the truth of generics is based on much looser *worldly truthmakers*. In this pa-

per, I develop a rearticulated account of the relationship between generics and the worldly truthmakers. The account is slightly different from Leslie's. Crucially, the *rearticulated psychological view* allows one to distinguish genuine generics like (3) from false generalizations like

(6) Pitbulls maul children.

(7) Muslims are terrorists.

Even though it is somewhat unclear, one interpretation is that Leslie thinks that (3), (6), and (7) are all generics. They are all generated by the psychological mechanism. (Leslie 2007, 384–385; 2017, 393–421.) In contrast, the rearticulated view allows one to distinguish a genuine generic from sentences that are false and therefore cannot be generics. After all, generics are, by definition, sentences that allow counterexamples but are still considered to be true. The rearticulation is based on the assumption that generics are a valuable source of information about the world. Only genuine generics convey valuable information, false generalizations not so much.

Moreover, the rearticulated psychological view enables a comprehensive response to Rachel Katharine Sterken's (2015) critical assessment of Leslie's psychological view. Sterken makes three claims: (i) Leslie's worldly truthmakers are open to numerous counterexamples; (ii) contrary to Leslie's thought, generics are context-sensitive; and (iii) generics do not express cognitively primitive generalizations. At the heart of my response is the distinction between genuine generics like (3) and sentences like (6) and (7) which admittedly do look like generics but in actuality are not generics at all. Namely, they are not supported by worldly truthmakers.

2. Semantic views

The semantic views have two important features. First, it is widely accepted among the semanticists that the structure of (1) is

Gen(x) [Tiger(x)] [striped(x)].

Secondly, the views rely on an extensional interpretation of the *Gen*-operator. To put it a bit crudely, the basic idea is that the *Gen*-operator specifies a relationship between two sets in

the scope of the operator. Furthermore, *Gen* is an operator over individuals. In (1), the operator picks the relevant proportion of individuals at the intersection of the set of tigers and the set of striped things. Importantly, the assumption is that the operator does the picking in a compositional way. To illustrate, existential and universal quantifiers are compositional too. They contribute to the truth conditions of sentences in which they appear in a systematic way. The truth conditions for

(8) Some *K*s are *F*.

require that the intersection of the set of *K*s and the set of *F*s is not empty. That is, at least one *K* has to be *F*. The truth conditions for

(9) All *K*s are *F*.

require that the set of *K*s is a subset of *F*s. Similarly, the semantic views aim to come up with a semantic interpretation for *Gen* so that it systematically picks the right proportion of individuals in the scope of the operator.

3. Against semantic views

3.1 Structure of generics

Leslie agrees that the structure of a generic like "*K*s are *F*" is

Gen(*x*) [*K*(*x*)] [*F*(*x*)].

That is, she agrees that generics involve a hidden operator over individuals. In contrast, David Liebesman proposes what might be called *kind-predication* according to which the structure of (1) is simply

(10) *Panthera Tigris* is striped,

in which *Panthera Tigris* is a noun phrase denoting a kind. Hence, generics involve predications of properties to kinds. (Liebesman 2011, 409–442.) Against this, consider:

(11) Cats lick themselves.

If (11) was a kind-predication, then any cat that licks another cat would make it true, but that is not what (11) means. So the following structure for (11) is much more plausible:

Gen(x) [Cat(x)] [Licks(x, x)].

This structure captures the idea that cats lick themselves. This tips the scales in favour of *Gen*-analysis, according to Leslie (2015, 34–39). Nevertheless, Leslie refrains from any further semantic analysis of *Gen*. She has two reasons for this.

3.2 Asymmetry in complexity

Leslie's master argument is what she calls *asymmetry in complexity*. The asymmetry in complexity is based on linguistic evidence concerning language-learning, especially in children. The studies concerning language learning suggest that children "find generics so much easier to comprehend than quantified statements [...]" while "[e]xplicit quantifiers, whose semantics have proved quite tractable for the theorist, are *more challenging* for the young child than generics." (Leslie 2007, 380.) However, the semantic accounts of generics are far more complex than the formal representation of, say, universal quantification. The semantic accounts often involve very sophisticated formal semantics, but the linguistic evidence suggests that generics are very easy to understand. Leslie concludes that there must be another explanation for generics that does not rely on highly sophisticated formal semantics. Nevertheless, I am not entirely convinced by this argument. It seems to me that linguistic competence and the formal representation of that competence are two different things, and a complex representation of an utterance does not mean that the utterance itself is difficult to understand. Consider the following sentence:

(12) Riding a bike without a helmet is dangerous.

I would assume that even small children understand (12) and, reluctantly perhaps, accept it as true. However, the formal representation of (12) is surprisingly complex. First, it is not a conjunction

(13) Riding a bike is dangerous and not wearing a helmet is dangerous.

because neither conjunct is true. Rather, (12) involves a compositionally complex expression of cycling and not wearing a helmet of type $\langle e \rightarrow t \rangle$, which is formed with lambda abstraction:

$\lambda x\ (Cy(x) \wedge \neg We(x))$.

The (second-level) property of being dangerous of type $\langle \langle e \rightarrow t \rangle \rightarrow t \rangle$ is then predicated on this complex (first-level) property:

$Da(\lambda x\ (Cy(x) \wedge \neg We(x)))$

Even though the formal representation of (12) is fairly complex, small children can still understand it and know it to be true. To me, this shows that linguistic competence and the formal representation of that competence are two different things.

3.3 Conjunctive generics

I think that the second reason involving *conjunctive generics* is more important than Leslie's master argument. Consider two generics:

(14) Peacocks lay eggs.

(15) Peacocks have fabulous tails.

Then form the conjunction

(16) Peacocks lay eggs and have fabulous tails.

People assent to this conjunction. However, this conjunction is very difficult for any extensional semantic view because (16) is not true of any single peacock (females lay the eggs and males have the tails). On the basis of this, Leslie says that, rather than based on extensional semantics, the inference from (14) and (15) to (16) is based on inferential rules. Specifically, (16) is based on the conjunction introduction rule. (Leslie 2007, 390–391 and 400.)

It is important to note that Leslie's explanation based on inferential rules is directly at odds with the semantic explana-

tion because the inferential rules like the introduction rule for conjunction

($\wedge$-*I*) $A; B \vdash A \wedge B$

holds regardless of the content of *A* and *B*. In this sense, you might call the inferential rules "logical." The aim of the semantic views is that the truth of a conjunction like (14) stems from the semantic content of (14) and (15), and, according to the semantic view, the *Gen*-operator has a crucial role in determining the content of (14) and (15) and therefore the truth conditions of (16). (This point is revisited in Section 6.2.)

4. Psychological generalizations and worldly truthmakers

In the face of the two problems, Leslie proposes a different approach. She says that generics are based on a primitive mechanism of generalization. Importantly,

> [t]hese cognitively primitive generalizations do not operate on set extensions, or any such abstraction. They are not grounded in such extensional or statistical information, but rather depend on factors such as how striking and important the information in question happens to be. (Leslie 2007, 394.)

According to Leslie, generics do not have truth conditions in the sense that the *Gen*-operator would contribute to the truth conditions compositionally. Rather, generics have much looser worldly truthmakers. Leslie distinguishes three types of generics: (i) majority generics, (ii) characteristic generics, and (iii) striking-feature generics. The main purpose here is to clarify the relationship between generics and the worldly truthmakers. While rearticulating is needed for all three types, the focus is on the striking-feature generics. It is the most interesting type and also the most controversial.

Concerning majority generics, the truth of (1) requires that the majority of tigers actually are striped. The world has to be such that the majority of tigers are striped. If it was that only a small number of tigers were striped, (1) would not be true.

(2) is a characteristic generic. This type of generalization categorizes kinds, such as animal kinds, on the basis of characteristic features, and reproduction is a characteristic feature.

It characterizes ducks as egg-layers. In contrast, (4) is not true because being a female does not characterize ducks in any significant way. Concerning the truthmakers for characteristic generics, the world has to be such that the male ducks are only negative counterexamples. That is, the male ducks do not present an alternative way of reproduction. They do not, for example, give birth to live ducklings. At the same time, (5) is also false because it makes a claim about the reproduction of bees. Hence, it should be construed as a characteristic generalization instead of a majority generalization. This is further discussed below but, at this point, it should be pointed out that the non-sterile bees are positive counterexamples which falsify (5).

Finally, (3) is a striking-feature generic. Leslie argues that the primitive mechanism of generalization is often triggered by information that is striking, horrific or appalling. The primitive mechanism is triggered because the mechanism is looking for a *good predictor* of the striking or horrific feature. Even though only a few members of the kind possesses the generalized property, one would still be well-served to be forewarned about the property. The truth of (3) relies heavily on the disposition to carry the virus: "It is important, for example, that the virus-free mosquitoes be capable of carrying the virus" (Leslie 2007, 385). That is, even if only a small portion of mosquitoes carry the virus, the rest are disposed to carry it.

At this point, two things should be mentioned. First, the constraint concerning the positive and negative counterexamples also applies to striking-feature generics. The second point is that the striking-feature generics are also a reason to favor the psychological view. Namely, the semantic views struggle to explain the truth of striking-feature generics like (3) just because less than 1% of mosquitoes actually carry the virus.

5. Rearticulating the psychological view

My central argument is that Leslie's view on the relationship between generics and worldly truthmakers needs rearticulation. The rearticulation comprises two things: (a) If we articulate the relationship between generics and the worldly

truthmakers more carefully, the articulation yields a clearer distinction between genuine generics and false generalizations. It seems to me that Leslie is not clear enough on this matter. (b) I argue that the psychological mechanism is optimized to our perceptual capacity. This can explain some of the puzzling aspects of generics. As we move on to Sterken's objections, all of these rearticulated items are discussed. As it turns out, (a) and (b) are crucial to my response to Sterken.

5.1 Distinction between generics and generalizations

To start with the distinction between genuine generics and false generalizations, Leslie gives the following examples of striking-feature generics:

(17) a. Mosquitoes carry the West Nile Virus.

b. Sharks attack bathers.

c. Pitbulls maul children.

Soon after this, Leslie adds the most controversial sentence to the list of striking-feature generalizations:

(18) Muslims are terrorists.

One rather plausible interpretation is that Leslie thinks that all of these sentences are generics. (Leslie 2007, 384–385.) They are all generated by the primitive mechanism of generalization. It is just that some of them are supported by the worldly truthmakers and others are not. (17a) is true while (18) is clearly false. In contrast, I propose a different view. I argue that only (17a) is a genuine generic and (17b), (17c), and (18) are not. The reason is that while they are no doubt products of the generalization mechanism, they are not supported by worldly truthmakers (more detailed reasoning below.) This emphasizes the role of worldly truthmakers in distinguishing genuine generics from those which are not. This is based on the assumption that generics are a valuable source of information about the world. False generalizations do not convey valuable information about the world. Leslie forms a worldly truthmaker constraint for a generic "*K*s are *F*":

The counterinstances, if any, are negative, and:

If *F* lies along a characteristic dimension for the *K*s, then some *K*s are *F*.

If *F* is striking, then some *K*s are *F* and the others are disposed to be *F*.

Otherwise, the majority of *K*s are *F*.

The rearticulation just insists that we stay faithful to these truthmakers. For example, if a bunch of bigots believe that all Muslims are disposed to commit terrorist attacks, that does not make (18) a generic. What is needed is that the world actually is such that all Muslims are disposed to commit terrorist attacks. But they are not. Hence, (18) is not a generic. Similarly, it is highly unlikely to me that pitbulls are disposed to attack humans or children specifically. Admittedly, there are some statistics that seem to support this idea. However, as an owner of a pitbull, I am aware of the problems that these statistics present. First, in many statistics pitbulls are categorized by type, not by breed. As a result, many crossbreed dogs are entered in the pitbull-type category. If, for example, a labrador-pitbull crossbreed bites someone, it is categorized as a pitbull, not as a labrador. This prejudices the categorization immensely. In fact, I cannot help thinking about the infamous and racist one-drop rule of the yearly 20th century legal system in the US. Furthermore, even if it is true that pitbulls do actually cause more problems than other dog breeds, it is most likely to do with the abuse they have endured as it is a fact that pitbulls are popular dogs in the cruel dogfighting business and among other abusers. So, on the basis of the statistics, you cannot tell if pitbulls have an inherent disposition to be aggressive towards humans or that other dog breeds or types lack this disposition. Therefore, there is no truthmakers for (17c). As we move on to (17b), my confidence fades a bit as I do not have a pet shark. However, there lies the problem, people generally do not have sharks as pets and they remain rather mysterious animals. Nevertheless, biologists who work on shark do seem to suggest that shark are more likely to swim away when they encounter humans. So it is more likely to be true that sharks are disposed to swim away when encountering a human being. Hence, even

though I am slightly uncertain about this one, I am inclined to rule (17b) as false.

My rearticulated view seems to be in conflict with Leslie's view. The different truth values in (17) shows this. However, it is difficult to tell where exactly we disagree. It could be that we disagree about the theory but it could also be that we disagree about the empirical facts concerning pitbulls and sharks. It seem to me that Leslie is somewhat vague concerning the distinction between genuine generics and false generalizations. She explicitly says that the sentences in (17) are true generics but it is somewhat unclear what she thinks about (18). She does not explicitly say if the fact that it is a (false) generalization triggered by the psychological mechanism is enough to make it a generic. If we exclude (18), then the difference between my rearticulation and Leslie's view might not be theoretical but rather a factual difference. It could be that Leslie and I simply disagree about the facts concerning pitbulls and sharks. Either way, I argue that this clarification between genuine generics and false generalization is crucial for the plausibility of the psychological view as discussed in Section 6.3.

5.2 Perceptual optimization

I claim that the best way to interpret the psychological mechanism is that it works in conjunction with our perception, and it is designed to be as efficient as possible (given our imperfect perceptual capacity). In other words, the generalization mechanism is optimized to our actual perceptual capacity. (3) illustrates this again. Given our poor ability to distinguish those mosquitoes which actually carry the virus from those which do not, the mechanism is locked on to the whole mosquito kind. If the virus made the mosquitoes grow ten times bigger and turned them bright orange, we would not have a generic like (3). Instead, we would have a universally quantified sentence

> (19) All huge and bright orange mosquitoes carry the West Nile virus.

Needless to say, this would be very convenient concerning the threat of the West Nile virus. But in reality, we cannot

identify the virus-carrying mosquitoes. So the mechanism is locked on to the entire kind. At the same time, we do have the capacity to distinguish mosquitoes from other insects. Thus, the mechanism is locked on only to mosquitoes, not to insects in general. It might be counterproductive to believe that insects carry the virus as that would cause needless panic. (Leslie 2007, 383–386.)

On the other hand, we could imagine that our perceptual capacities were much better than they actually are. Imagine that we could smell viruses just like some dogs can smell some viruses. Let's assume that the odor of the West Nile virus resembles vanilla. We then could have a universally quantified sentence:

(20) All mosquitoes with a hint of vanilla scent carry the West Nile virus.

Leslie herself does not talk about this aspect of the psychological mechanism, but it seems to me that this addition is very much in line with what Leslie says about the purpose of the mechanism:

> It is clear that this mechanism ought to be an efficient information gathering mechanism, since it is our most basic and immediate means of obtaining information about categories. One way such a mechanism might be efficient is for it to take advantage of regularities out there in the world. (Leslie 2007, 383–384)

If the mechanism is tuned to its highest efficiency, then surely it should accommodate our imperfect information gathering mechanisms—in this case, our inability to distinguish virus-carrying mosquitoes from virus-free mosquitoes either visually or by the odor.

6. Sterken's three objections

The rearticulated relationship between generics and the worldly truthmakers has an important role in my response to Sterken's objections. She argues that (i) Leslie's worldly truthmakers are open to numerous counterexamples; (ii) contrary to Leslie, generics are context-sensitive; and (iii) generics do not express cognitively primitive generalizations. (i) is

divided to two: counterexamples to characteristic generics and counterexamples to striking-feature generics. While responding to characteristic counterexamples, we also get a response to (ii), which questions Leslie's claim that the *Gen*-operator does not have a compositional contribution. The objection is based on Sterken's claim that only a semantic interpretation of *Gen* can explain the context-sensitivity associated with generics. Finally, even though my response to (i) and (ii) at least partly relies on my rearticulated view, the response to (iii) relies solely on my rearticulation.

A few points about Sterken's strategy should be mentioned as her strategy also affects my counterstrategy. First, concerning the counterexamples, Sterken claims that because there are *numerous* counterexamples to Leslie's view, the evidence just keeps stacking up against Leslie. I go on to demonstrate that this thought is erroneous. There are not *numerous* counterexamples to the psychological view. Secondly, her discussion focusses on striking-feature generics. Namely, she argues that there are no striking-feature generics. If this was the case, it would indeed be a severe blow to Leslie's view because striking-feature generics are the most celebrated feature of her view. Striking-feature generics set the psychological view apart from the other views because they can explain why (17a) is a genuine generic. If it turns out that there are no striking-feature generics, then Leslie "loses a great deal of the evidence for her psychologically based theory—plausibly the best evidence for a psychologically based theory," says Sterken (2015, 2503). Her denial of striking-feature generics leads to the third point. She argues that we should not always trust our intuition about generics. Even though the sentences in (17) seem like genuine generics, it turns out that they are not. I agree that we should not always trust our intuition about generics. As I already mentioned, I agree that (17b) and (17c) are not generics. However, this does not mean there are no striking-feature generics at all because (17a) is such. There are striking-feature generics, but they are not as common as Leslie thinks.

6.1 Counterexamples to characteristic generics

The counterexamples to characteristic generics have to be negative. Sterken presents the following set of characteristic generics. She says that all of them have positive counterexamples:

(21) a. Mammals give birth to live young.

b. Birds fly.

c. Swedes have blond hair.

d. Dutch people are tall.

e. Reptiles lay eggs.

f. Dobermans have floppy ears.

Sterken argues that these are all genuine characteristic generics, but they have positive counterexamples.

At the very beginning, I admit that (21a) is a genuine counterexample to Leslie's view and also to my rearticulated view. The platypus is a positive counterexample to (21a): Platypuses are mammals but they lay eggs. However, the rest are not counterexamples to the psychological view. That means that the evidence does not stack up against the psychological view. There are not numerous counterexamples to the psychological view.

To start with (21b), I do not think that is a characteristic generic. It is a majority generic. Sterken says that there are about 40 species of birds that cannot fly. Given that there are over 18 000 species of birds, (21b) well passes muster for a majority generic. Importantly, with majority generics, it does not matter whether the counterexamples are positive or negative. The plausibility of the rearticulated view then turns on the question about the nature of the ability to fly. Is it a characteristic feature or just a feature that the majority of birds share? One of the key features of the rearticulation is the order of truthmakers. According to the rearticulation, characteristic or striking-feature generalization trumps majority generalization. In our present case, it seems to be somewhat problematic. On the present interpretation of (21b), the ability to fly is a feature that the majority of birds have. Initially, one

might think this is counterintuitive. Surely, it is characteristic for birds that they fly. However, to maintain the rearticulated psychological view, I have to insist that the ability to fly is not a characteristic feature of birds. The seeming characteristic nature of flying among birds stems from the fact that the vast majority of birds do fly.

It is true that the stereotypical conception of the Swedes is that they are blonds. But that is only a stereotype, and stereotypes are very often misleading. As Sterken says, contrary to the stereotype, many Swedes have brown hair. (Sterken 2015, 2497.) (Just because stereotypes can be misleading, I will not go through Sterken's points which are based on stereotypes. I think this is justified given that the examples based on stereotypes have only a minor role in her argumentation.)

It is important to distinguish (21c) from (21d): (21c) is based on a stereotype, but (21d) is based on a fact. The average height of the Dutch is the tallest in Europe. As such, (21d) does seem to present a tricky case for the psychological view. As Sterken points out, every short Dutch person is a positive counterexample to (21d). However, we should be clear whether the counterexamples are against the generic sentence or against its truthmaker. I argue that it is against the truthmaker. In that case, we need to be clear about the truthmaker. Concerning 21d, the truthmaker is the statistical fact that the Dutch are the tallest in Europe *on average*. This has a dramatic effect on the counterexamples. The shorter Dutch are no longer counterexamples to the truthmaker. The shorter Dutch people are included in the average height of the Dutch. Rather, a counterexample would be a taller average height in another European country. But there is no such counterexample. It is a fact that, on average, the Dutch are the tallest in Europe. After Greg Carlson, it could be argued that (21d) should be interpreted as

(22) The average Dutch person is tall.

According to Carlson, (22) is a genuine generic but the interpretation of the noun phrase is purely intensional. The term "average Dutch person" does not have an extension. You cannot have lunch with the average Dutch person. (Carlson 1989, 167–192, especially 184.) Nevertheless, this is not what we are after here. My counterargument rests solely on the

distinction between generics and their truthmakers. According to the current proposal, (21d) is interpreted as "*Gen x* [Dutch(*x*)] [tall(*x*)]" and (22) is the truthmaker for the generic. In other words, (21d) is true because it is a statistical fact that, on average, the Dutch are the tallest in Europe. As result, if (21d) is interpreted as a characteristic generic, as Sterken intended, it is true because it has no counterexamples. It is also true, if it is interpreted as a majority generic because it is a statistical fact that the Dutch are the tallest in Europe.

Initially, I thought that (21e) is true, but when Sterken laid out the facts about reptiles, I changed my mind. Namely, there are plenty of reptiles that give birth to live babies: snakes, chameleons, and some lizards (Sterken 2015, 2497). In the light of this evidence, I am ruling (21e) as false and so it cannot be a genuine generic. Here we can see one important consequence of my rearticulated view. In order to evaluate which sentences are genuine generics and which are not, you need information about the world. Knowledge about worldly truthmakers is very important when figuring out genuine generics. In some cases, one must go against one's initial urge to generalize certain features across the whole kind. Especially, if there is contrary evidence as (21e) illustrates.

Concerning (21f), Sterken's informants thought that it is true. However, in my informal inquiries, the most common answer was something like "Erm, don't they have pointy ears?" So, according my informants, (21f) is false and a better candidate for a generic might be

(23) Dobermans have pointy ears.

If indeed (21f) is a generic at all, you might view it as a majority generic. A quick picture search revealed that, in the first 30 pictures, a Doberman had pointy ears in 25 pictures. So it is very typical that Dobermans have pointy ears. Sterken does admit that the truth of (21f) requires a very specific context:

> [(21f)] uttered in a context in which the speaker is discussing the biological properties of dobermans, is intuitively true despite the fact that most dobermans have the alternative property of possessing pointy ears. (Sterken 2015, 2497.)

On the basis of this, Sterken argues that generics manifest context-sensitivity. (21f) is true when talking about the bio-

logical properties of Dobermans (and (23) is false). (21f) is false when discussing dog breeders' aesthetic standards (and (23) is true). This point is at the heart of Sterken's objection (ii), as she thinks the context-sensitivity of generics is strong evidence for a semantic interpretation of the *Gen*-operator. Sterken proposes a test for context-sensitivity of generics:

> *A-quantifier test*: Substitute the (hidden) *Gen*-operator with explicit adverbial substitutes like "typically" or "normally." If there is no variation in the truth conditional contribution between the explicit substituents, then these cannot be the source of contextual variation. Therefore, the source of contextual sensitivity has to be *Gen*.

If this test is applied to (21f), we then have two versions:

(24) a. Typically, dobermans have floppy ears.

b. Normally, dobermans have floppy ears.

Sterken relies on her informants again. She reports that her informants think that both of them are false regardless of the context. It would seem that this rules out the usual adverbial suspects and the culprit for contextual sensitivity has to be *Gen* since the generic form is the only one that presents context-sensitivity. According to Sterken, this is bad news for Leslie because she does not give any semantic interpretation of *Gen*. As Sterken aptly points out, context-sensitivity could easily be explained with quantificational domain restriction, but this requires an extensional treatment of *Gen* which Leslie refuses to give. (Sterken 2015, 2503–2505.) I assume that, with the quantificational domain restriction, Sterken means a situation in which the domain from which the *Gen*-operator picks up the relevant individuals is contextually restricted. For example, when Oxford University announces that all students are required to report to the vice chancellor's office by the end of week, it does mean that every student in the world needs to report to the office, just the Oxford students. The domain in this case is restricted to Oxford University, even though it isn't explicitly said in the announcement. Similarly, you could say that, in (17c), the domain is restricted to adult Dobermans and, in (23), the domain is restricted to Doberman puppies. Crucially, the restriction relies on a strict analogy

between the semantic interpretation of the universal quantifier and the semantic interpretation of the *Gen*-operator.

However, Sterken's test is far from conclusive. Without hesitation, I say that (24a) is false and I would imagine my informants would say that too, given their belief that Dobermans have pointy ears. But I hesitate with (24b). My intuition says that it is true that normally, without any interference, Dobermans do have floppy ears. So it is far from clear that the only possible culprit for context-sensitivity is *Gen*. There are other suspects for it. I think this is enough to cast a doubt on the idea that there has to be an extensional interpretation for *Gen*. To be clear, I am not taking a stand on the question of whether generics are context-sensitive or not. All I am saying is that if they are, then Sterken has not shown a reason to think that the responsibility for the sensitivity rests solely on *Gen*.[1]

---

[1] While discussing the context-sensitivity of generics, Sterken offers another point against Leslie. Sterken argues that in the following examples, the a-sentences are false generalizations, but when they are contextually embedded in b-sentences they become true:

1. a. Mammals lay eggs.
   b. Birds lay eggs. Mammals lay eggs, too.
2. a. Novels are paperbacks.
   b. Manuscripts are always paperbacks. Novels are paperbacks, too.
3. a. Bees are sterile.
   b. Many insects face reproductive challenges. However, only bees are sterile.

Nevertheless, I do not think these examples are successful. Let's consider a publishing editor encountering sentences like (1b)–(3b). I would imagine that she would have a lot to say about them, namely that, as they stand, they are either highly misleading or downright false and they need rewriting:

1* b. Birds lay eggs. Some mammals lay eggs, too.

2* b. Manuscripts are always paperbacks. Many novels are paperbacks, too.

Finally, 3b is clearly false because bees are not sterile. So it needs considerable rewriting:

3* b. Many insects face reproductive challenges. However, only bees are on the brink of sterility.

6.2 Counterexamples to striking-feature generics

Striking-feature generics are a crucial part of Leslie's psychological view, as the view handles nicely generics like (renumbered here as)

(25) a. Mosquitoes carry the West Nile Virus.

b. Sharks attack bathers.

c. Pitbulls maul children.

Other views struggle to explain these. However, according to Sterken, this turns out to be false advertisement. The celebrated feature of Leslie's view should not be celebrated because there are no striking-feature generics. As a consequence, there is no need to explain them. In contrast, I argue that there are striking-feature generics but fewer than Leslie thinks. According to my rearticulated view, only (25a) is a genuine generic, the others are not.

Sterken starts with the truthmakers for striking-feature generics:

> "*K*s are *F*" is true if:
>
> (i) the counterinstances (if any) are negative and;
>
> (ii) if *F* is striking, then some *K*s are *F* and the others are disposed to be *F*.

She points out that, according to this disposition clause, as she calls it, many false generalizations come out as striking-feature generics. For example,

(26) Humans kill themselves.

Suicide is a pretty striking and horrific and, notably, only humans commit suicide. Sterken also says that the counterexamples, those humans who do not kill themselves, are negative. On the face of these facts, it seems that in Leslie's view (26) is a genuine generic. Yet in reality, it is false.[2] Sterken

[2] Interestingly, some of my informants thought that "Humans commit suicide" is true. They thought that it is true because only humans commit suicides. However, I am not confident enough to say that this should be the sole objection to Sterken. Still, I think it is worth mentioning.

grants the possibility that (26) has not zoomed in to the right predictor of suicide: "Perhaps amongst humans, there is a subclass which serves as a better predictor" (Sterken 2015, 2501). So let us consider:

(27) Depressed people kill themselves.

Here the predictor zooms in to a set of depressed people but still (27) is false, according to Sterken. When responding to this, it should be remembered that the mechanism latches on to the whole mosquito-kind due to perceptual optimization. We cannot distinguish between the mosquitoes which carry the virus from the mosquitoes which do not. So, by locking on to the entire mosquito-kind, the mechanism is as efficient as it can be. However, with suicide we can do better than (27). Namely, we can consult various medical professionals. They could inform us that severe depression coupled with, say, *XYZ*-disorder is a high risk factor in committing suicide. Consider,

(28) Severely depressed people with *XYZ*-disorder commit suicide.

This might be true but, in my view, there are genuine moral reasons not to put it this way. The mechanism generalizes striking and often negative features but Leslie and others have argued that the mechanism can also work the other way round. The generic form can lead to generalizing and *essentializing* negative features of a social kind. This again leads to a negative view of that social kind because it is thought that the negative feature is an essential feature of the kind (with no possibility of a cure).[3] (Rhodes et al. 2012, 1–6.) The important point is the contrast with (25a). Even if we consulted the experts in the field of mosquitoes, we still would not be able to distinguish virus-free mosquitoes from those which carry it. To repeat, the mechanism locks on to the best possible predictor, given our imperfect perceptual capacity.

Nevertheless, Sterken's final blow to Leslie's view is that this (or any manoeuvre like this) cannot save the psychologi-

[3] I am much more comfortable with a phrase like "Severe depression coupled with XYZ disorder is a high risk factor in suicide."

cal view because the damage is already done with (25a). The real problem is that the disposition condition for the mosquitoes is already too weak. "To get a sense of just how weak the disposition clause must be," Sterken invites us to consider:

(29) Insects carry the West Nile virus.

According to her, this is intuitively true. But since the disposition to carry the virus is locked on to mosquitoes, (29) comes out false in Leslie's view. (Sterken 2015, 2501.) But here is my question: In what sense is (29) true? Generics famously do not confirm to any obvious monotonicity patterns, and the contrast between (25a) and (29) is a vivid example of this. Since only about 1% of mosquitoes carry the virus, it is tempting to say that (25a) presents similar monotonicity patterns as the existential quantifier. But that would be a mistake. Existential quantification is an upward monotonic quantifier. Namely, you can always go from the subset to the superset:

(30) Some tall men like tea. Thus, some men like tea.[4]

So if the hidden *Gen* in (19a) was similar to the existential quantifier, then the inference from the subset of mosquitoes to the superset of insects would be good. However, there is a strong negative response to (29), something like "Not all insects!" So to say that (29) is true is highly counterintuitive. Indeed, it is part of the appeal of the psychological view that it can explain why a generic like (25a) does not conform to the monotonicity patterns. The mechanism is locked on to mosquitoes, not to insects in general. Sterken actually captures the explanation perfectly: "[O]n a strict reading of Leslie's disposition clause [(29) is] false since not all insects share the relevant disposition of carrying disease [...]"[5] (Sterken 2015, 2501).

---

[4] In contrast, the universal quantifier is downward monotonic:
All men like tea. Thus, all tall men like tea.

[5] Sterken's further example also turns against herself. She thinks that, according to Leslie's view
4. Homosexuals carry HIV.
is true but, in reality, it is just a prejudicial and false generalization (Sterken 2015, 2502). But why would it be true in Leslie's view? The disposition to carry HIV is not limited to gay people. Heterosexuals are dis-

It should also be emphasized that monotonicity is a semantic notion. Hence, the weird monotonicity patterns support Leslie's claim that the only acceptable inferences involving generics are based on rules of inferences as inferential rules are not semantic. As discussed earlier, the conjunctive generics are not based on any extensional interpretation of the *Gen*-operator, according to Leslie. Rather, they are based on "logical rules" like the conjunction introduction rule. I called them "logical" because the introduction of conjunction of *A* and *B* is independent of the semantic content of *A* and *B*. In the present context, it can argued that the inferences concerning generics are not based on the usual extensional monotonicity patterns. Rather, they are based on perceptual optimization. The generalization is locked on to mosquitoes because mosquitoes are the optimal kind in relation our capacity to distinguish one insect kind from another.

6.3 Generics and primitive generalizations

Sterken's final claim is that generics are not based on primitive generalizations. The most compelling evidence for this is disagreements concerning striking-feature generics. Consider the following disagreements:

(31) *A*: Let's stay inside. Mosquitoes are out there, and they carry the West Nile virus.

B: That's not true. Almost none of them do.

(32) *A*: Pitbulls maul children.

B: That's not true. There have only been a few isolated incidences.

(33) *A*: Sharks attack bathers.

B: That's not true. They almost never do.

Sterken argues that these are all genuine disagreements. Furthermore, *B*'s responses are quite compelling. So the disa-

---

posed to carry it too. So the mechanism is not locked on to homosexuals and (4) comes out false in Leslie's view.

greements from (31) to (32) show that the striking-feature generics are systematically false:

> These kinds of dialogues I suggest should be taken as evidence that [the sentences in (25)] are not true in general—when we think they are true we are making a mistake. (Sterken 2015, 2010)

According to Sterken, this also suggests that there are no striking-feature generics since all of them are false. Nevertheless, here are similar disagreements. Only *B*'s responses are changed. Importantly, *B*'s altered responses reflect the central feature of my rearticulated psychological view:

(31*) *A*: Let's stay inside. Mosquitoes are out there and they carry the West Nile.

B: That's true. Fortunately, only few actually carry it. Unfortunately, we cannot tell which ones.

(32*) *A*: Pitbulls maul children.

B: That's not true. Various studies show that pitbulls are no more dangerous than golden retrievers.

(33*) *A*: Sharks attack bathers.

B: That's not true. Only around 0.00...002% of bathing instances involve shark attacks.

Admittedly, *B* is a very well-informed participant. She has extensive knowledge of mosquitoes, the probabilities of shark attacks, and studies on pitbulls. I think this reflects the fact that it takes a bit of knowledge to separate genuine generics from those which are not. In my rearticulated psychological view, we need the knowledge about the worldly truthmakers because the worldly truthmakers only support genuine generics. As it turns out, the worldly truthmakers support only (25a). In contrast, Sterken thinks that even the sentence about mosquitoes is false. From this, she infers that there are no striking-feature generics. In contrast, I argue that there are striking-feature generics. However, there are fewer of them than Leslie thought. Only generics supported by the worldly truthmakers are genuine generics and the generic about mosquitoes is supported by worldly truthmakers.

7. Conclusion

The central feature of the rearticulated psychological view is the insistence that the worldly truthmakers should be taken seriously. This enables one to separate genuine generics from those which are not. This is particularly important concerning striking-feature generics. It significantly narrows down the number of striking-feature generics. Still, according to the rearticulated view, there are striking-feature generics.

I have shown that Sterken's claims from (i) to (iii) are far from conclusive. The fact there are fewer striking-feature generics than Leslie thought does not mean that there are no striking-feature generics at all, as Sterken suggests. Moreover, there is no conclusive argument from context-sensitivity that the *Gen*-operator has to be interpreted compositionally. Finally, I have countered the claim that there are *numerous* counterexamples to the psychological view. There is only one counterexample, that pesky Platypus. This counterexample could be downplayed in various ways. For example, it would probably turn out to be a very challenging case for any view of generics, but I will not argue for that here. Instead, I admit that it is a real counterexample even to the rearticulated view and it deserves more attention. However, I think that that is beyond the scope of this article.

Tampere University

References

Carlson, Greg, N. (1989). "On the Semantic Composition of English Semantic." In G. Chierchia, B. H. Partee, and R. Turner (eds.), *Properties, Types and Meaning, Volume II: Semantic Issues*. Dordrecht: Kluwer Academic Publishers, 167–192.

Leslie, Sarah-Jane (2007). "Generics and the Structure of the Mind." *Philosophical Perspectives* 21, 375–403.

Leslie, Sarah-Jane (2008). "Generics: Cognition and Acquisition." *Philosophical Review* 117(1), 1–47.

Leslie, Sarah-Jane (2015). "Generics Oversimplified." *Noûs* 49(1), 28–54.

Leslie, Sarah-Jane (2017). "The Original Sin of Cognition: Fear, Prejudice and generalization." *Journal of Philosophy* 114(8), 393–421.

Liebesman, David (2011). "Simple Generics." *Noûs* 45(3), 409–442.
Rhodes, Marjorie, Sarah-Jane Leslie, and Christina M. Tworek (2012). "Cultural Transmission of Social Essentialism." *Proceedings of the National Academy of Sciences (PNAS)*, 109(34), 1–6.
Sterken, Rachel Katharine (2015). "Leslie on Generics." *Philosophical Studies* 172, 2493–2512.

The Dual Character of Essentially Contested Concepts

JOONAS PENNANEN

1. Introduction

This paper puts forward and examines the claim that essentially contested concepts (hereafter ECCs)—as they are originally presented by W.B. Gallie in his seminal paper "Essentially Contested Concepts" (Gallie 1956b)—share a conceptual structure with dual character concepts (hereafter DCCs) first identified by Joshua Knobe, Sandeep Prasada, and George Newman in "Dual Character Concepts and the Normative Dimension of Conceptual Representation" (Knobe, Prasada, and Newman 2013). The proper employment of ECCs is said to inevitably involve endless and rationally irresolvable yet genuine disputes that are sustained by perfectly respectable arguments and evidence. DCCs are concepts that encode both a descriptive dimension and an independent normative dimension: people employing DCCs have been found to be employing two sets of criteria of category membership that match with the two dimensions, which makes it possible to judge a given object as a category member in either or both senses.

I do not seek to show that ECCs and DCCs match one-to-one with each other. Instead, I explore their distinct and theoretically significant structural affinities that make way for a better understanding of these concept types and their structures. I argue that ECCs encode a descriptive and a normative dimension in much the same way as DCCs. This connection may be thought as accidental or as a mere similarity that does not justify further conclusions, however, and that is why I further bolster my case by juxtaposing natural kind concepts (hereafter NKCs) with ECCs and DCCs. Concepts are particu-

larly elusive objects of study. By a three-way comparison I seek a firmer ground for the identification of genuine similarities that indicate a shared structure, as surprising as the combination of these concept types may seem at first. I show that making categorizations with DCCs and NKCs requires a reference to an underlying deep structure, and I argue that it is also the case with ECCs. This ultimately means that psychological essentialism has an important role to play in the phenomenon of essential contestability.

Much of my argument rests on evidence amassed by comparing different perspectives on concepts, and therefore it is best to note in advance that both DCCs (see Knobe, Prasada, and Newman 2013; Newman and Knobe 2019) and ECCs (see Evnine 2014) have been directly linked to NKCs before. However, no such connection has been proposed as holding between DCCs and ECCs until this paper. At the end of the day, I claim that the structural commonalities between these three types of concepts outweigh their respective differences for the purpose of explaining the nature of ECCs, specifically. By no means do I wish to suggest that all questions one may have about ECCs will be answered, or even can be answered, by this account. Instead, I hope to offer a theoretical framework for seeing ECCs in a new light and for understanding why many of the issues arise in the first place, especially regarding alleged essentialist underpinnings of Gallie's thesis. Structural similarities between mostly theorized ECCs, recently identified DCCs, and the already well-established class of natural kinds should make ECCs less mysterious as objects of study. Exploring the shared conceptual characteristics should also offer further guidance on which conceptual operations are possible in the case of each concept type, but apart from a few general suggestions made here and there, I am content to leave it to future research.

2. Of essentially contested concepts, accrediting valued achievement, and contestation

At the heart of Gallie's account is a claim that is both striking and unnerving: "there are concepts which are essentially contested, concepts the proper use of which inevitably involves endless disputes about their proper uses on the part of their

users" (Gallie 1956b, 169). Gallie seeks to show that these disputes are genuine and "sustained by perfectly respectable arguments and evidence" even if they are not "resolvable by argument of any kind" (ibid.). There are only four concepts that are originally deemed essentially contested by Gallie: ART, DEMOCRACY, SOCIAL JUSTICE, and CHRISTIANITY.[1] In a later revised work, SCIENCE is included as well, though with some reservations (Gallie 1964, 156, 190). Despite its influence in various fields (see Pennanen 2021, sec. 2.6), Gallie's thesis in its original form is unclearly articulated and highly controversial. Subsequent theorists have typically tried to reconstruct the thesis after which they have discussed and dissected what they understand as its cardinal claims, merits, and failings.[2] A systematic or adaptive reconstruction is beyond the scope of the present paper (instead, see Pennanen 2021), but we should still start by presenting the most important characteristics of ECCs as Gallie understands them.

Gallie offers us seven conditions for ECCs (hereafter "Condition(s)" with Roman numerals as presented below), yet he refers to them as the "conditions of essential contestedness" as well.[3] The Conditions are:

---

[1] Gallie uses several different terms and phrasings interchangeably, i.e., "religion" (Gallie 1956b, 187; 1964, 168), "the adherence to, or participation in, a particular religion," "a Christian life" (ibid., 180; 1964, 168–69), "the Christian tradition," and "Christian doctrine" (ibid., 168; 1964, 157). In his final formulation, Gallie appears to prefer CHRISTIANITY (Gallie 1964, 168–70). For a further discussion, see Pennanen 2021, 57, n. 52, 179–84, 451, 462–64. Throughout the text, I will use small capitals to name and refer to concepts.

[2] For a comprehensive overview of various positions, see Collier, Hidalgo, and Maciuceanu 2006; Pennanen 2021.

[3] In Gallie's original texts, the phenomenon of interest is named as "essential contestedness." In literature, it is often presumed that a correct or at least philosophically interesting form is "essential contestability." In the same vein, "essentially contested concept" is often replaced with "essentially contestable concept." These are not interchangeable; for a discussion, see Pennanen 2021, sec. 12.2, 12.3. In the current paper, however, I will disregard this complication as far as the terminology is concerned and refer only to "essentially contested concepts," or ECCs. "Essential contestability" is reserved for a general phenomenon, and "essential contestedness" is invoked only in the case of Gallie's original thesis.

> *Condition I:* The concept must be "appraisive in the sense that it signifies or accredits some kind of valued achievement." For example, many would urge that democracy "has steadily established itself as the appraisive political concept *par excellence*."
>
> *Condition II:* "This achievement must be of an internally complex character, for all that its worth is attributed to it as a whole."
>
> *Condition III:* "Any explanation of its worth must therefore include reference to the respective contributions of its various parts or features; yet prior to experimentation there is nothing absurd or contradictory in any one of a number of possible rival descriptions of its total worth, one such description setting its component parts or features in one order of importance, a second setting them in a second order, and so on." Therefore, "the accredited achievement is *initially* variously describable."
>
> *Condition IV:* "The accredited achievement must be of a kind that admits of considerable modification in the light of changing circumstances (…) the concept of any such achievement [is] "open" in character." Later, Gallie asserts Condition (IV) to state "that the achievement our concept accredits is persistently vague."
>
> *Condition V:* "[E]ach party recognizes the fact that its own use of it is contested by those of other parties, and that each party must have at least some appreciation of the different criteria in the light of which the other parties claim to be applying the concept in question."
>
> *Condition VI:* "[T]he derivation of any such concept from an original exemplar whose authority is acknowledged by all the contestant users of the concept."
>
> *Condition VII:* "[T]he claim that the continuous competition for acknowledgement as between the contestant users of the concept, enables the original exemplar's achievement to be sustained and/or developed in optimum fashion." (Gallie 1956b, 170–173, 180, 182)

Gallie's Conditions have attracted a lot of criticism and most commentators have ended up eschewing one or more of them for various reasons (see Collier, Hidalgo, and Maciuceanu 2006; Pennanen 2021). The orthodox interpretation of Gallie's

thesis locates the endlessness and inevitability of disputes in the characteristics of a concept which render the disputes over the uses of that concept endless and incapable of being rationally settled (see, e.g., Swanton 1985, 813–15; Bryant 1992, 58; see also Gallie 1956b, 188). Yet it has been argued that, for someone genuinely holding an essential contestability view, there is no sense in engaging in a contest which cannot by its nature be won or lost (Gray 1983, 96; Zimmerling 2005, 25; see also Connolly 1993, 226; but cf. Swanton 1985, 815; Waldron 1994, 534). Gallie himself did not rule out the possibility of temporary agreement for practical reasons (cf. Gallie 1964, 211). This arguably leaves room for genuine disputes even if the critical points raised are found to be basically sound (Pennanen 2021, sec. 13.1–13.2). All in all, it is far from a trivial matter where exactly to draw a line between such a dispute's conceptual, practical, and substantive elements, but since it does not directly pertain to the structure of concepts, we can note this and move on.[4]

A chief theoretical worry with respect to ECCs is the possibility that no independently plausible theory of concepts will be able to allow a type of conceptual structure that admits endless and rationally irresolvable disagreements over one and the same and/or mutually shared concept as Gallie claims (Gallie 1956b, see 169, 188, 190, 196; but see also 1964, 177, 211). For instance, according to Frege's view in *Grundgesetze der Arithmetik*, the definition of a concept must be complete, and it must unambiguously determine whether a given object falls under the concept or not; concepts that are not sharply defined cannot be recognized by logic (Ricciardi 2001, 52ff). More generally, especially among philosophers there is a widely held assumption that properties, propositions, and relations that are candidates for being members of linguistic expression are precise in that a number of objects either definitely instantiate or definitely fail to instantiate them; any proposition is likewise either definitely true or definitely false (Braun and Sider 2007, 134). More simply, semantic objects that are designated by concepts or linguistic

[4] In this article, I do not examine the sense in which relevant disputes are endless and irresolvable either. For a review of a variety of positions, see Pennanen 2021, in particular sec. 12.4.

meanings are thought to be precise. That general standpoint is assumed by the classical theory of concepts which holds that concepts have a definitional structure, i.e., they encode necessary and sufficient conditions for their own application (Laurence and Margolis 1999, 8–9). Clearly, or so it may be claimed, there cannot be genuine contestation over the kind of concept that is understood to pick its object(s) precisely or without any ambiguity or underdetermination. Making conflicting claims that presumably originate in different and quite possibly equally reasonable uses of the same concept is thus ruled out by logical fiat. Yet there is an even simpler way of understanding the problem of conceptual unity, and it generalizes beyond the classical view: how can mutually contesting ways of concept employment serve as legitimate uses of one and the same concept despite the alleged differences at a conceptual level, differences that are meant to generate a dispute in the first place? Relevant differences would also mark different concepts (see also Newey 2001).

The aim of the current paper is not to address the issue of conceptual confusion. Neither do I focus on values or principles or the substance of concepts; an essential contestability thesis is about "structures and procedures" (Freeden 2004, 7). But what do the structures and procedures cover? The last three of Gallie's Conditions belong to pragmatics rather than to semantics (van der Burg 2017; Pennanen 2021, chap. 10), and if all seven Conditions are understood as conditions of *a concept*, an ECC seems to involve more than is typically understood to fall under a concept's structure. That is why I am introducing the notion of a *conceptual architecture*, within which I am including the pragmatic circumstances or the context in which people characteristically employ a concept as well as that which is semantically encoded in the concept. The distinction between a concept's structure and its architecture is not completely clear-cut; for example, a concept's relations to other concepts can reasonably fall in either category—choosing this way or that way ultimately depends on one's favored theory of concepts. Neither is my terminological choice completely innocuous: it allows me to discuss the normativity of concept employment without taking a stand on whether that normativity is primarily located in concepts *qua* concepts or in the ways they are employed (i.e.,

in particular contexts). As a result, I do not take a stand on whether contestability is a feature of (certain) concepts or their context of employment. The possibility of conceptual confusion, or the unity problem, is something that I cannot avoid discussing in the following even if I do not claim to provide a solution to it.[5]

To compare a normative dimension of DCCs to that of ECCs, I first need to say a bit more about the way normativity figures in the conceptual architecture of ECCs. On the face of it, the first three Conditions are the most relevant, yet one also needs to pay attention to Gallie's general approximation of what his thesis is about. Gallie does not unambiguously explicate what he means by the notion of "appraisiveness," yet it is clear that his focus is on positive appraisal, i.e., something is taken as an achievement and is evaluated favorably (cf. Gallie 1956b, 184). This positive appraisal is then coupled with a standard that is mutually recognized in spite of the dispute (Gallie 1956b, 197; see also Weitz 1972, 103–4). Gallie's reference to achievements looks to be quite literal: if parties to a dispute consider a thing an achievement, they certainly evaluate it favorably. Contested concepts "pick out activities, practices, or goals that the community's members are prepared to praise in others or strive to achieve themselves" (Criley 2007, 33). According to this notion, ECCs should be understood as normative—it is reasonable to further specify the relevant sense as evaluative as there are standards of evaluation involved (van der Burg 2017, 234; cf. Gallie 1956b, 197). Still, Gallie's choice to go with "appraisive" instead of "evaluative" may also be taken to

[5] All these questions cannot be discussed in just one paper. Nevertheless, I should note that the talk of "architecture" instead of "structure" at this juncture is partly motivated by my doubt that a specific feature of generating endless and irresolvable disputes about a concept's proper use could be encoded in some singular concepts as their invariant and stable feature. For an argument to this effect, see, e.g., Newey 2001, and see Pennanen 2021 for full discussion of the unity problem. In addition, I will briefly summarize central features of the essential contestability thesis that I deem defensible in footnote 19 in sec. 5.

indicate that he refers to normative assessment and judgment more generally.[6]

With the introduction of Conditions (II) and (III), we learn that the achievement in question is meant to be internally complex and variously describable. The idea is that the complex parts or features of the valued achievement are all understood to contribute to what makes the achievement worthy of admiration. By arguing for their views, disputing parties are understood to be advancing different descriptions of the valued achievement, descriptions in which the component parts or features are differently ranked. Therefore, when ECCs become contested, it makes sense to think that there are diverging personal or group-specific evaluations or preferences, which result in conflicting descriptions of the correct way of using the concept, and a mutually recognized standard (of evaluation) at work at the same time. That which is mutually recognized by the disputants appears to have a role of bringing some unity to contestation, yet Gallie clearly thinks that it cannot serve as "a general principle" that decides the issue once and for all (cf. Gallie 1956b, 177–79, 189).

Gallie approximates the way ECCs are contested by presenting an artificial scenario in which different teams vie to be the champions in a continuously proceeding game. A championship in this game is awarded on very unusual grounds: the team that gathers the most support or followers is (effectively) dubbed the champions. Spectators support their chosen teams based on who plays the game best, or the

[6] Much of the scholarly work done in relation to Gallie's original thesis has revolved around interpreting what he means, or reconstructing what he should mean, by ECCs being "appraisive." For different interpretations, see, e.g., Weitz 1972, 103–4; Gellner 1974, 95; Gray 1978, 392; Connolly 1993, 10, 22–3; Freeden 1996, 55–56; Lukes 2005, 14; Collier, Hidalgo, and Maciuceanu 2006, 237; Criley 2007, 33; Boromisza-Habashi 2010, 277; Väyrynen 2014, esp. 472, 474–8, 487; van der Burg 2017, 233–34, n. 16. Some view Gallie's focus on a positive appraisal as an unfortunate mistake; they claim that there is really no reason to omit unfavorable evaluations from the scope of essential contestability (Freeden 1996, 55–56; see also Garver 1987, 220; Collier, Hidalgo, and Maciuceanu 2006, 216). This is correct if one's aim is to assess concepts that figure in all sorts of normative judgments, but Gallie's original writings do not support that interpretation (Pennanen 2021, sec. 4.1, 11.3).

way the game is meant to be played, and each team comes to be ranked based on the level of their specialized or otherwise distinct way of playing the game. Gallie fleshes out the example by describing one particular game that resembles bowling. He observes that

> such bowling can be judged, from the point of view of method, strategy and style, in a number of different ways: particular importance may be attached to speed or to direction or to height or to swerve or spin. But no one can bowl *simply* with speed, or simply with good direction or simply with height or swerve or spin: *some* importance, however slight, must, in practice, be attached to each of these factors, for all that the supporters of one team will speak of its "sheer-speed attack" (apparently neglecting other factors), while supporters of other teams coin phrases to emphasise other factors in bowling upon which their favoured team concentrates its efforts. (Gallie 1956b, 173)

Different ways of bowling that are attached with importance represent, outside the artificial example, various aspects or features of a valued achievement that can be ranked differently. It is important to recognize that in both his Conditions (namely II, III, and V) and the description of the artificial example, Gallie requires concept-users to hold the same descriptive features as at least somewhat important aspects of the valued achievement. Contestant teams compete "for the acceptance of (what each side and its supporters take to be) the proper criteria of championship" (Gallie 1956b, 171). As there are "no official judges or strict rules of adjudication" (ibid.) that would decide the question of which team is the most deserving of the championship, the game can go on even after determining the level of support each team has at any given time. In other words, supporters of *every* contesting team continue to regard their favored team as "the champions" or perhaps as "the *true* champions," "*morally* the champions" etc. (ibid.) unless they are convinced otherwise. So even if all groups of supporters may acknowledge the effectiveness of one team in gathering the most supporters, "the property of being acknowledged effective champions carries with it no universal recognition of outstanding excellence—in [a team's] style and calibre of play" (ibid.). The above translates to continuous contestation by concept-users about how

to properly rank various aspects or features of a valued achievement. But, of course, the artificial example is meant to serve as a ladder to Gallie's theoretical claim about ECCs: the proper uses of these concepts are persistently contestable and actually contested by others. To the extent that these concepts have a standard or general usage, it consists of mutually contesting and mutually contested uses (ibid., 169).

The artificial example ends with Gallie affirming that the supporters "continue with their efforts to convert others to their view, not through any vulgar wish to be the majority party, but because they believe their favoured team is *playing the game best*" (Gallie 1956b, 171). I think it is safe to say that the artificial groups of spectators/supporters and contesting teams are meant to coalesce into one in real life. We are evaluators who (passively) deem things better or worse, and agents who (actively) seek to advance or bring into effect that which we consider valuable. A big part of the latter are our attempts to persuade our fellow men. This is enough of Gallie's thesis for now. I will continue examining the nature of ECCs after first taking a look at DCCs and NKCs.

3. Of dual character concepts and natural kind concepts

DCCs are concepts that encode both a descriptive dimension and an independent normative dimension (Reuter 2019, 1). Concept-users have been found to be employing two sets of criteria for category membership that match with the two dimensions, which makes it possible to judge a given object as a category member in either or both senses (Knobe, Prasada, and Newman 2013, 243, 246–49, 253–54). More specifically, there are cases in which concept-users think that an object is clearly "X" but is not "true X," or is not "X" but is "true X," or is both "X" and "true X." This "double dissociation" sets DCCs apart from a more common notion that category membership can come in degrees (ibid., 253).[7] Dual character concepts have been distinguished by testing, for instance, how a person responds to statements that have a particular form such as "there is a sense in which she is clearly not a scientist, but ultimately, if you think about what it really means to be a

[7] For early seminal views on the notion of graded membership, see Lakoff 1973, Rosch and Mervis 1975, and Hampton 1979.

scientist, you would have to say that she truly is a scientist" (ibid., 242). In some test scenarios, participants make up their minds with the help of vignettes that provide them with additional information regarding, for instance, the said scientist's motives, capabilities, *et cetera*. Another method is to assess how sensible given statements are when a key term is changed. Based on their experiments, Knobe, Prasada, and Newman conclude that DCCs "support two types of normative judgments ("good" and "true") whereas the control concepts support only one of these types of normative judgment ("good")" (ibid., 245; see also Newman and Knobe 2019; Liao, Meskin, and Knobe 2020).[8]

DCCs have a specific organization or structure that sets them apart from most other concepts. They are "represented via both (a) a set of concrete features and (b) some underlying abstract value" (Knobe, Prasada, and Newman 2013, 243). A given set of concrete features will cohere "because they are all ways of realizing the same abstract values" (ibid., 256), and so the two sets of criteria for the application of a DCC can both be derived from the same set of concrete features. Regarding criteria that match the descriptive dimension, concept-users simply check whether a given object has the right features. In the case of criteria that match the normative dimension, concept-users identify the abstract values that the concrete features serve to realize and then check to see whether the object in question displays these values (ibid., 254). The structure of DCCs can be further elaborated, some-

[8] The list of DCCs that are tested by Knobe *et al.* 2013 includes FRIEND, CRIMINAL, LOVE, MENTOR, COMEDIAN, MINISTER, THEORY, BOYFRIEND, ARTIST, ARGUMENT, TEACHER, POEM, SOLDIER, SCULPTURE, ART MUSEUM, MUSICIAN, MOTHER, ROCK MUSIC, SCIENTIST, NOVEL. The control concepts are MECHANIC, OPTICIAN, BAKER, BLOG, DOORMAN, MAYOR, WAITRESS, CASEWORKER, TABLE OF CONTENTS, TAILOR, BARTENDER, RUSTLING, WELDER, CATALOG, CHAIR, FIREFIGHTER, UNCLE, CASHIER, STROLLER, OBITUARY, SECOND COUSIN. In the experiment that involves the judgments "good" and "true" (one of the total five) participants were instructed to rate the sentences "That is a good *x*" and "That is a true *x*" with DCCs and control concepts substituted with "*x*" as to how natural or weird they sounded.

what surprisingly, by comparing them to the category of natural kinds.[9]

The natural kind terms refer rigidly to things in the world: the real determinant of the extension is a natural property. The indicators of a concept are thus contingent in that they only point toward an underlying natural essence; the underlying reality provides one with the final criteria (or norms, rules etc.) that constitute the concept (or govern the intension of the respective term). To illustrate, "is wet" may be taken as an indicator that one might be dealing with a natural kind "water," yet water's underlying essence is H_2O. The fact that water is wet is an observable feature of the natural kind "water" but there is a clear sense in which it is merely superficial as far as categorizing items accurately as water is concerned. Even if we would be inclined to think that water in steam form is wet, a solid block of ice certainly is not until it melts. The contingency of indicators is perhaps even more obvious in the case of species categories. Tigers may very well be striped and ferocious but that is neither a necessary nor sufficient criterion for their category membership as tigers. Instead, there is an underlying causal factor (a tiger's hidden essence if you will) that is ultimately decisive.

Knobe, Prasada, and Newman contend that the same structure is at work with both DCCs and NKCs: "In both cases, people show a willingness to go beyond concrete observable features, and in both cases, they seem to be understanding categories in more abstract theoretical terms" (Knobe, Prasada, and Newman 2013, 254). How this plays out with NKCs is clear enough. With DCCs, like ROCK MUSIC or MOTHER, people associate the concept "with a collection of features, but they then face a further question about why the category is associated with those specific features and not others" (ibid., 255). The criteria governing the concept give an answer to this question yet, "this time, the answer is not that all of the features share the same underlying causes but rather that they all embody the same abstract values" (ibid.). This is arguably a significant difference: the order of concrete observa-

[9] A concise yet useful characterization of natural kinds is provided by Crispin Wright in "The Conceivability of Naturalism" (Wright 2003, 359–60) which is the one that I have made use of in this paper.

ble features and an underlying understanding of a category is reversed. Whereas the features or indicators that are typically enumerated for NKCs are brought about by an underlying essence, in DCCs the features directly contribute to the realization of some same abstract value(s), i.e., they bring the value that underlies the category about. The question becomes: do the structural similarities between NKCs and DCCs outweigh the differences?

There are further studies that show that the distance between NKCs and DCCs is, at first blush, not as great as one might think. First, Newman and Knobe (2019) draw attention to a body of evidence that suggests that people tend to represent some concepts in terms of a deeper unobservable property or "essence." Although most of the research on such psychological essentialism has so far been focused on patterns of judgment found for natural kind concepts such as TIGER or WATER, essentialism plays an important role in many other cases as well.[10] Of special interest presently are socially constructed concepts that are ordinarily understood to invoke certain values or ideals (or, they are regarded as "value-laden"). Newman and Knobe claim that these concepts—of which they specifically mention SCIENTIST, CHRISTIAN, and ART (ibid., 586; see also Liao, Meskin, and Knobe 2020; compare Gallie's list of ECCs in sec. 1)—reflect the same underlying cognitive structure that is applicable in the case of NKCs: the tendency to try to explain observable features in terms of a further unifying principle. With NKCs, one is dealing with causal essentialism: "the essence of a natural kind is understood as the underlying cause of its various superficial features" (ibid., 587). In the case of socially constructed concepts, essentialism is "Platonic," i.e., "people appear to believe that what binds together the different features of the category is the fact that they are all ways of embodying the same deeper value" (ibid., 588). Nevertheless, both are cases of (psychological) essentialist representation: there is an unobservable

[10] "Psychological essentialism" was first dubbed as such by Medin and Orton 1989; see also Medin 1989. For more references to studies on both psychological essentialism and more specifically on the (ordinary speakers') use of natural kind terms, see Newman and Knobe 2019 and Haukioja, Nyqyist, and Jylkkä 2021, 378–81.

property that is responsible for category membership, and that binds a concept's superficial features together (ibid., 589).

Second, Tobia, Newman, and Knobe (2020) have conducted a series of experiments[11] that aim to uncover people's actual intuitions about Hilary Putnam's famous Twin Earth thought experiment as far as categorization of Twin Earth "water" is concerned. In the thought experiment, Twin Earth "water" has the same appearance, taste(lessness), and other apparent qualities and functions (e.g., it is clear, quenches thirst, and supports life) as Earth water in normal conditions, yet Twin Earth "water" has a complex chemical formula abbreviated as XYZ that essentially differs from H_20.[12] Instead of endorsing or rejecting what Tobia *et al.* take as the standard philosophical intuition (cf. Haukioja, Nyquist, and Jylkkä 2021, 397), i.e., that the Twin Earth liquid is not water, research participants were found to assent to two distinct claims: (i) there is a sense in which the liquid is water; and (ii) ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid is not truly water at all (Tobia, Newman, and Knobe 2020, 183). In other words, test subjects' complex reactions to Twin Earth cases displayed a dual character pattern, which Tobia *et al.* take as evidence in favor of the view that NKCs are also employed by making use of two sets of criteria—one set is based on underlying causal properties, the other on superficial properties (ibid.).

Tobia, Newman, and Knobe do not claim to have settled the question of which theory of natural kind categorization process is correct, or what the final semantic implications of their findings might be. Research participants' judgments about category membership were found to depend on the context of categorization, which makes a range of interpretations possible (see ibid., 197–205). Nevertheless, they do state that any plausible theory about NKCs should be elaborated to *account for* the dual character pattern of judgment (esp. ibid., 203). The further claim that NKCs share a conceptual structure with DCCs is indirectly supported by recent studies that have either called into question the prevalence of the philo-

[11] For details, see Tobia, Newman, and Knobe 2020.
[12] For specifics, see Putnam 1975.

sophical Twin Earth intuition or have otherwise demonstrated that NKCs are represented both in reference to their underlying structure and superficial qualities or even by their appearance alone in certain cases (see Haukioja, Nyquist, and Jylkkä 2021).

As to the essentialist employment of DCCs and NKCs in categorization, I assume as an intermediate conclusion that they have similar conceptual architectures. The present extension of theoretical scope from concepts' structures to their architectures is meant to reflect also the finding that the context dependence of terms that denote[13] NKCs may be compatible with several ways of understanding and organizing their criteria of application. Furthermore, in practice, people seem to use terms that denote NKCs in a way that admits double dissociation which is the hallmark of DCCs. This implies that both concept types have structures that consist of two distinct criteria for categorization. Given that the specific aim of my examination is to pump insights to better explain ECCs and the disputes in which they are involved, there is no need to show that NKCs and DCCs have exactly the same conceptual structure, not to mention broader architectures—previous observation about the difference between causal essentialism and Platonic essentialism is more than enough to show that this is not the case. From the standpoint of ECCs and the disputes in which they are involved, what ultimately matters is that categorization judgments are made in like manner with DCCs and NKCs in practice, or at least can be made. In the next two sections, I will argue that the mutual similarities extend also to ECCs. For this, we need to pick up the discussion where it was left at the end of section one.

4. Examining the architecture of essentially contested concepts in light of dual character concepts

The valued achievement signified by an ECC is understood as internally complex, which results in the conception that there are multiple criteria by which an object may fall under the banner of a concept. Disputing parties endorse conflicting

[13] In the current paper, I am using the word "denote" in its ordinary meanings "to serve as an indication of" and "to stand for."

descriptions of the appropriate way of employing the concept; this involves diverging personal or group-specific evaluations in addition to some mutually recognized standard or background that unifies otherwise centrifugal evaluative disagreement (see also sec. 1). Given the centrality of evaluation in the description of how ECCs are characteristically employed aggressively and defensively in a dispute, it is reasonable to assume that to satisfy the application criteria of ECCs "is to satisfy a norm of excellence, as well as a mere precondition of a classification" (Gellner 1974, 95; see also Gray 1978, 389). These criteria play a dual role: they are criteria according to which one evaluates the worth of the achievement itself (the norm of excellence part) but they can also be viewed as the criteria that need to be met for an object to be judged as falling under the concept (the classification part) (see also Pennanen 2021, 388). As both description and evaluation are needed for employing ECCs aggressively and defensively in a dispute (cf. Condition V in sec. 1), this may lead one to conclude that ECCs are neither purely descriptive nor purely evaluative (see, e.g., van der Burg 2017, 233–34). A dispute over ECCs is best understood as conceptual and substantive (Gray 1978, 391), or as conceptual, normative, and substantive (Besson 2005, 16, 71–72) depending on emphasis.

In the artificial example, contestation takes place over which factor, or which weighted combination of factors, is the most important for playing the game best. Different ways of playing contribute to overall excellence in the game. In formal terms, one should note that Condition (II) has two parts: one stating that an achievement signified by a concept is internally complex, i.e., it admits multiple descriptions; another stating that the worth of the achievement is attributed to it in its entirety. The value of the achievement, or the overall excellence of playing a game as it is meant to be played, is considered to be at least somewhat independent of available ways of employing the concept, or of the ways or styles of playing that game. This feature of ECCs' conceptual architecture is also shared by DDCs: there is a value or ideal that underlies a concept, and the features that are picked by the concept's descriptive criteria cohere just because they are all ways of realizing an abstract value (that is signified by a concept), and the

concept's appropriate use needs to meet with the value/ideal at least in certain circumstances.[14]

Contestation over ECCs follows when individuals or groups come to advocate for their own evaluation of which way of employing the concept meets with the underlying value or ideal best, but it appears to take place solely on the evaluative and not on the descriptive side. This is because Gallie, in effect, subscribes to the view that separates descriptive concepts (or conceptual elements) from evaluative concepts (or conceptual elements). The former are "responsive to the co-presence of a number of distinct descriptive or naturalistic features of the world, each of which must be of equal weight" while the latter are not "flatly conjunctive" but "can be responsive to these descriptive or naturalistic features in a way that reflects different weight or influence among the descriptive features" (Criley 2007, 36). This enables the users of ECCs to argue that although all proposed alternatives may be, for instance, democracies in some relatively clear sense, only one of them is worthy of being called a democracy. This type of judgment is absolutely central to essential contestability and closely resembles the double dissociation that is the hallmark of DCCs.

The descriptive and normative dimensions of ECCs and DCCs are both similarly independent, but there is also reason to think that categorizations made by employing these concepts involve the same type of normative judgment. A recent discussion of social role DCCs—certain social role concepts such as SCIENTIST or ARTIST are sometimes taken to be the pa-

[14] As it is, the value-ladenness of ECCs, or essential contestability in general, has been understood in the literature in terms of the inescapability of normative perspective (Connolly 1993, 10, 22–23), as disputes between rival moral and political commitments and/or perspectives (Lukes 1977, 418–19; Gray 1978, 392; Grafstein 1988, 19, 25), or as caused by employing a concept that is oriented towards an ideal which allows endless debate about precisely what it implies (van der Burg 2017, 233–234). Moreover, it has been argued that it is part of the meaning and function of some words "to indicate that a value judgment is required" (Waldron 1994, 527) or that the rule for the correct use of certain contestable concepts is "to elicit a specific value judgement from anyone applying or implementing the proposition in which they appear (Besson 2005, 82). I will briefly mention yet another formulation of value-ladenness by Stokes 2007 in sec. 5.

radigmatic examples of DCCs (Del Pinal and Reuter 2017, 477; see also Leslie 2015; Del Pinal and Reuter 2015)—is very helpful for clarifying the matter. Not all social role concepts are DCCs (e.g., WELDER, BUS DRIVER) though. The normative dimension of social role concepts that obtain higher ratings as DCCs from participants in experiments may have only little to do with the usual or typical function of the corresponding social roles. Instead, the normative dimension of DCCs represents more like an idealization of the basic function of the role (Leslie 2015; Del Pinal and Reuter 2017). For example, being a "true parent" is not solely about having offspring but also involves caring deeply and supporting one's ward.

This type of idealization is also what Gallie had in mind. To see why this is the case, let me first note that Gallie views RELIGION or CHRISTIANITY as the concept that best satisfies the seven Conditions of essential contestedness (Gallie 1956b, 180–81). In his later *Philosophy and Historical Understanding*, he emphasizes that he wants to consider CHRISTIANITY "in its practical, not its purely doctrinal, manifestations e.g. as exemplified by what would generally be meant by such a phrase as 'a Christian life'" (Gallie 1964, 169). The account that immediately follows only partially connects with social roles, yet near the end of "Essentially Contested Concepts," Gallie notes that

> *Some* of our moral appraisals command universal assent, but by no means all do so. It is of the first importance to insist that we also use the word "good" (or its near-equivalents and derivatives) with a definitely moral, but just as definitely questionable force: witness such phrases as "a good Christian", "a good patriot", "a good democrat", "a good painter" (when we mean a sincere, sensitive, intelligent, always rewarding—but not necessarily a "great" or a "fine" painter), "a good husband," and so on. In all these uses, it seems perfectly clear, our concept of the activity in and through which the man's goodness is said to be manifested, is of an essentially contested character. (Gallie 1956b, 195)

Gallie's general idea is that in the case of above social roles, it seems always possible to contest what it really means to be a good husband, for instance, by proposing different criteria for it. What is new is that Gallie now draws attention to the ex-

pression "a good painter" hoping to clarify a special sense that differs from a comparatively unexceptional matter of evaluative degree that "great" or "fine painter" more accurately indicates. Elsewhere, Gallie expresses that sense by using the modifiers "true" (ibid., 171, 177, 178) and "more orthodox" (ibid., 177) which corresponds nicely with the way DCCs are characteristically employed. Even without a comparison to DCCs, it is quite clear that Gallie's idea of essential contestedness is premised upon the possibility of idealizing (and/or interpreting) in different ways that which is considered to be of value in the case of certain activities or achievements. Literature on DDCs simply clarifies the issue by presenting less complicated examples of the type of normative judgment that is also present with ECCs. Other concepts do not admit such idealization. A welder can certainly be regarded as good at welding, yet (most) people do not find it sensible to speak of "true welders" while the expression "a true artist" is sensible in (most) normal contexts. In "Art as an Essentially Contested Concept" (Gallie 1956a), Gallie speaks of the contestability involved in determining what should count as "a work of art" but he could have just as easily said "a true work of art," the once-and-for-all uncontestable determination of which requires lasting agreement on what art truly is.[15]

Although there has not been much discussion of conceptual contestability in the literature on DCCs, unlike ECCs, they draw attention to disputes that arise from conflicting descriptive and evaluative uses of a concept.[16] Disputants now employ somewhat distinct sets of criteria that are distinguished by their type rather than employing different sets of criteria of the same type but with more or less different content. Both

[15] I will continue drawing examples from this article in the two remaining sections as well in order to better connect this type of judgment with other elements of Gallie's thesis.

[16] The possibility of using words both descriptively and evaluatively does not escape Gallie. According to him, the history of art "discloses a growing recognition of the fact that the word 'art' is most usefully employed, not as a descriptive term standing for certain indicatable properties, but as an appraisive term accrediting a certain kind of achievement" (Gallie 1956a, 111). Nevertheless, his argument is not about this type of contestability.

cases may be taken as confusions or non-genuine conceptual disagreements but for different reasons. In the former case, if parties are sharing the same concept along with its two sets of criteria of application, their disagreement can easily be resolved by pointing out that they are just employing different sets of criteria. There is no real disagreement unless one of the parties has made some kind of mistake in applying the concept, or disagreement is factual in that parties do not agree on which features an object, to which the concept is applied, has. In the latter case, a parsimonious explanation of what is going on looks to be that the different sets of the application criteria mark different concepts, and not just different uses or functions (perhaps distinguished by type of criteria) of presumably one and the same concept. Gallie himself effectively dismisses the possibility that the relevant type of contestedness originates in a contest over which features should be ranked in the first place (cf. Gallie 1956b, 174, n. 2). Relegating contestation strictly on the evaluative side aims to avoid a situation in which people are simply talking past each other by underpinning the unified identity of conflicting concept-uses to mutually accepted descriptive criteria. Unfortunately, this may result in a sense of contestability that is somewhat impoverished or not far-reaching enough.[17]

In the same vein, analyzing ECCs and DCCs side by side raises the question of whether it is possible for DCCs to be essentially contested. All it would seemingly take is that an abstract value that underlies a concept, and by virtue of which concrete features cohere, were to be contested by disagreeing parties. That would be problematic for reasons that are instructive more generally. To share a concept opposing

[17] For reasons why a farther-reaching or more encompassing contestability deserves the appellation "essential contestability," see Pennanen 2021, sec. 15.3. It is not uncommon to claim that essential contestability requires something more or beyond normative disagreement or the absence of universally agreed schemes of values (see Freeden 1996, 55). For instance, Peter Ingram views some concepts as *partially* contestable in that they can be *evaluatively*, but not essentially, *contested*. Evaluative contestation is made possible by the fact that certain concepts "necessarily possess certain, agreed common features" or properties while the essential contestability proper becomes more a matter of family resemblance-type fluidness of criteria (Ingram 1985, 44–45).

parties need to accept certain things about it and the dispute over the concept('s application) needs to pertain to other things. In the case of DCCs, a category's "essence" can be understood as a placeholder with a clear function: it brings together the features of that category/concept as ways of realizing an abstract value. A dispute over that which unifies the concept, even if cashed as an abstract value or ideal, questions the unity and raises the uncomfortable possibility that parties to the dispute are just talking past each other. The situation is no different in the case of ECCs assuming that they are structurally similar enough to DCCs as the current examination into both concepts' conceptual architectures suggests: disagreement on a ranking order may be taken as evidence of a disagreement that is ultimately about an ECC's deep structure, or about a value or ideal that provides the rationale for grouping certain features together. If this is correct, and contesting such a rationale opens the door for contesting descriptive criteria as well, it is ultimately the reason why one cannot hope to guarantee the unity of the concept by insulating descriptive criteria from contestation. Once the genie is out of the bottle, essential contestability looks to persistently threaten the sense in which concept-users are sharing the same concept. Understanding ECCs as dual character concepts makes no difference based on the current analysis.

The conceptual operations that have been discussed should not necessarily be viewed as mutually available to the other concept type, and nor should it necessarily be thought that, for instance, an essentially contested DCC metamorphoses into an ECC when a value that serves as its deep structure is mutually contested. A typology according to which double dissociation and essential contestability are defining features of ECCs and DCCs is also an option: when concepts whose criteria of application play a dual role in dividing between descriptive and normative criteria of application, and which refer to a deep structure that underlies categorizations, are employed to "doubly dissociate," we are dealing with a concept having a dual character; when concepts' employment results in endless and irresolvable disputes, we are dealing with an ECC. That way there is room for not only ECCs that are not DCCs, and vice versa, but also to different interpretations of these phenomena (or Gallie's original thesis, for that

matter). Not much of substance hinges on this choice though, and it is likely going to be decided not only based on operative theories but also one's scholarly aims.

5. Examining the architecture of essentially contested concepts in light of natural kind concepts

To complement the picture of ECCs' specific conceptual structure, I now turn to discuss the conceptual architecture of ECCs in the light of NKCs. Natural kinds and that which is represented by ECCs may appear too different to engender fruitful comparisons at the level of terms and concepts, but Simon Evnine succeeds in finding important commonalities between them. Evnine claims that natural kind terms and essentially contested terms [sic] are both species of a single semantic genus: both types of terms "are, on the respective theories, correctly applied to something now if and only if it bears a certain kind of relation to samples or exemplars that have played an historical role in the use of the term" (Evnine 2014, 127). In the case of natural kind terms, the exemplars are natural while the operative relation is *belonging to the same kind as*. Here, "something like a deep structure (…) is tacitly assumed to underlie the operative relation" (ibid., 129). In the case of essentially contested terms, the exemplars are cultural while the operative relation is *being the heir of*, a component of which is the relation of *being part of the same tradition as* (ibid., 130).

Evnine's interpretation of Gallie's thesis closely resembles the semantic externalist theory of reference as it is put forward by Putnam and Kripke (Evnine 2014, 126–27), and Evnine finds a lot of significance in Gallie's sixth Condition, i.e., that any ECC or a use of ECC is to be derived from an original exemplar whose authority is acknowledged by all the contestant users of the concept. Yet, the exemplars of NKCs also differ in important respects from those of ECCs:

> Natural kind terms are typically names of kinds of natural *objects* or *substances*— water, tin, tigers, electrons. And the exemplars themselves are either objects of the relevant kind or quantities of the relevant substance. In the case of essentially contested terms, the exemplar is something like a stage of a tradition. The exemplar will therefore consist in anything that

> might be an element of a tradition: cultural objects (e.g., literary works, codes of law), institutions, ways of doing things, people and their actions and intentions, and people's understandings of all of the above. Evnine 2014, 127

An essentially contested term has the function of picking out something that "has the relation of being the heir of that tradition-stage" (ibid., 130), i.e., of the exemplar. In addition, the internal complexity of the exemplar allows one to pick any element of the tradition and synecdochically treat it as an exemplar itself (ibid., 127–28). For instance, in the case of essentially contested CHRISTIANITY one may treat the Bible and/or the biography of Jesus Christ as an exemplar but the deeds of apostles, ritual practices, and even moral principles and habits of early twentieth century church-goers (and many things more) could also be picked by one's usage of "Christianity" as authoritative (see also ibid., 127–28).

The reference of essentially contested terms like "Christianity" and "art" is historically connected to an exemplary phenomenon. One employs such terms correctly when a referred-to thing has the relation of being the heir to that exemplary phenomenon. The correctness of specific uses may of course be contested. Think of many intense disputes that revolve around the question of who the true heir or successor in a given instance is—the conflict between the Sunni and Shia Muslims is often mentioned as an example. Moreover, Evnine says that "[t]he exemplary phenomena and the things to which such terms correctly refer through their relation to these exemplary phenomena, constitute historical traditions," and the kind of contests that Gallie talks about are, in a manner of speaking, over ownership of traditions (ibid., 119). Such contests are endemic to traditions rather than essential to some group of concepts, and therefore Evnine prefers to speak of essentially contested terms instead of concepts. The relation of heirship does not necessitate or even imply rival claimants even if "it is highly likely that groups will evolve that prioritize the elements of [an exemplar that is rich in internal complexity] differently and hence that a contest will emerge over which party is the real heir of the exemplar" (ibid., 125). This makes Evnine's interpretation a variant of *an*

admittance to a tradition thesis of essential contestability (see Pennanen 2021, 233 see also sec. 18.4).[18]

Evnine's reframing answers the question of why Gallie thinks that the clarification of a concept's status as essentially contested requires viewing the concept with a historian's eye in addition to laying out its general (or "logical") characteristics (Gallie 1956b, 181–82, 196–97; see also sec. 5 below). As to ART, one needs a historical account of how ART *came to be* which comes down to seeing how and why presumably equally competent people have favored different and even radically opposed aesthetic standpoints. This helps one to appreciate the peculiar structure of ART and to see that it belongs to concepts that are "*essentially complex,* and, chiefly for this reason, *essentially contested*" (Gallie 1956a, 107). According to Gallie, there are several "classic theories or definitions of art" or "main types of aesthetic theory": configurationist theories, theories of aesthetic contemplation and response, theories of art as expression, theories emphasizing traditional aims and standards, and communication theories. Each theory "has been a contestant for the title of the true, the only satisfying, the only plausible theory of art" and "[e]ach is still capable of exercising a certain pull on our sympathies" (ibid., 112). Nevertheless, as theories that exclude other reasonable aspects of art, they are "intelligible only as contributions to a seemingly endless, although at its best a creative, conflict" (Gallie 1964, 177). Evnine's account explains why such historical understanding is required. If a term's applications conflict but otherwise seem reasonable individually one may be dealing with an essentially contested term instead of a confusion. Given that essentially contested terms are correctly applied to something if and only if that something has the relation of being the heir of samples or exemplars that have had a role to play in the use of the term, determining the matter requires assessing whether conflicting applications are

[18] David-Hillel Ruben (2010; 2013) has presented a substantively similar interpretation that focuses on the notions of true succession and faithfulness (to the original exemplar) within a tradition. As Ruben concentrates mostly on (social) epistemological issues, I omit discussing it here. That said, I am indebted to him for considerably broadening my own perspective on Gallie's thesis and essential contestability in general.

traceable and faithful to past exemplars and samples, and whether, as such, they are intelligible.

It seems plausible that dual character terms are also species of the same semantic genus as terms denoting ECCs and NKCs given that DCCs and ECCs on the one hand, and DCCs and NKCs on the other hand, have been found to share important characteristics. This potentially opens new avenues for study; for example, concerning conceptual judgments in relation to terms that signify social roles. Terms are not quite the same thing as concepts, but I think that one can go a bit further concept-wise (see also sec. 5). The move to a conceptual level can be made explicitly, for instance, by following Newman's and Knobe's (2019) lead: even if essentialism comes in many forms, they argue that people's reasoning about NKCs such as TIGER and WATER and essentialist-like intuitions that include people's representation of socially constructed concepts [or DCCs] like SCIENTIST or CHRISTIAN reflect the same underlying cognitive structure. Assuming this is correct, one may reasonably conjecture that people's representation of ECCs also reflects the same cognitive structure—the sameness should be understood here as a suitably broad generalization or type instead of complete identicalness—and that conceptual operations that are typical to ECCs are not necessarily far off even in the case of NKCs.

However, there are also important differences between the conceptual architectures of ECCs and NKCs despite their commonalities. A historical connection between an exemplar and a term seems to admit much more contestability in the case of ECCs than NKCs. I do not think that it has to necessarily mean that the link between historical samples and exemplars and its current usage is any less social/causal per se, which might suggest a different semantic genus. Following Evnine's view that contests are endemic to traditions, it seems plausible that the difference is attributable to the fact that the communities of experts which ultimately determine the correct way of employing natural kind terms are simply missing or otherwise found lacking in the case of "essentially contested terms." When people come forth with competing (and possibly reasonable) definitions or descriptions of the achievement in question, the conceptual architecture of ECCs has historically been formed such that it simply allows more

room for different conceptions to gain traction and be established as reasonable alternatives (see also Pennanen 2021, 211–13). As long as deference to experts is part of the conceptual architecture of NKCs, or the conceptual and linguistic practice of employing terms that denote NKCs, there is little reason to suspect that endless and rationally irresolvable disputes are about to spring forth. And just as was the case with DCCs before, at some point we may deem such changes significant enough for a given term to denote a different type of concept altogether.

6. Further reflection and theoretical implications

In this final section, I (a) present how the kinship between ECCs, DCCs, and NKCs reflects on the nature of ECCs while I also (b) assess how it all fits with Gallie's original ideas. Furthermore, in anticipation of criticism, I briefly clarify (c) why I am not confusing empirical and conceptual or logical levels of analysis, and (d) why the sort of essentialism that I advocate is not pernicious.

What can we say about the conceptual architecture of ECCs based on previous findings? Both DCCs and NKCs entail a reference to a deep structure that binds together different features picked by a concept, and such "hidden essence" looks to be tacitly assumed also in the case of ECCs. Instead of "Causal essentialism," however, I argue that we are dealing with a modified form of "Platonic essentialism" (cf. sec. 2). An ECC is involved in a dispute when mutually contested and contesting uses of a concept (or even concepts[19]) are

[19] I do not think that much of substance would be lost by understanding ECCs as second-order concepts or categories of possibly distinct first-order concept uses (cf. esp. Gallie 1956b, 169 about mutually contesting and mutually contesting uses that *make up* some kind of concept). In this picture, essential contestability is primarily about what concepts people should form or adopt in the first place, and "essentially contested concept" may be best considered as a term of art. According to the full essential contestability thesis that I view as defensible, the relevant type of contestability is brought about by anthropocentric concept employment that aims to persuade others (see Pennanen 2021, sec. 18.5 esp.). In short, "ECCs" aim to make the best sense of not only the proper boundaries of (participatory) human activities but they also have the function of facili-

faithful to exemplars and samples, all of which belong or are claimed to belong to the same tradition (or one of its branches) on the grounds that they embody and/or manifest the same abstract value or normative ideal (compare with Newman and Knobe 2019, 588; Evnine 2014, 127–30). The tradition is now understood as an open-ended human activity or practice with a temporal continuity. This still requires some clarification.

In discussing essentially contested DEMOCRACY, Gallie asserts that he is not concerned with either (descriptive) "questions of actual practice" or those "theoretical considerations" that suggest that either democratic or undemocratic consequences flow from certain arrangements. Instead, these particular uses presuppose "a more elementary use" that expresses political aspirations which have been embodied in countless "revolts and revolutions as well as in scores of national constitutions and party platforms" (Gallie 1956b, 183–84; see also Pennanen 2021, sec. 11.3). In the current framework, such "elementary use" is understood to depend on the conceptualization that aims at the true representation of a historically embodied normative ideal or value, the aim that exhibits psychological essentialism. A sample that bears or manifests the normative ideal or value becomes a part of a tradition that is viewed as sustaining and advancing that ideal or value. When people differ on what realizes the ideal or value best, they also come to disagree on how the respective concept is to be applied (see also Besson 2005, 82–83) or, more fundamentally with respect to essential contestability, on how the concept should be formed in the first place (see Pennanen 2021). At stake is not a direction of some social movement per se but effecting changes in how people con-

---

tating the best possible solutions to basic human problem areas and/or in connection to broadly understood activities in thought and practice. It follows that contesting concept uses have an endorsement function in addition to an interpretive function, and their contestability is thus a contextual and functional rather than a structural matter which is brought about by the fact of our human condition and an always-present practical possibility of questioning what we should do and why. Most of these features still belong to a concept's architecture as it is understood in this paper even if explaining the origin of essential contestability requires a (separate) metaphysical thesis.

ceive of and conceptualize issues of importance. Such changes in outlooks may then lead to other changes in the world through concept-users' subsequent doings.

There is still arguably a disconnect between my earlier general characterization of ECCs, the comparisons of DCCs and NKCs to ECCs, and the present picture in historical terms. The dilemma is similar to the problem that Gallie too faces: it appears that our present understanding of certain ideals and values is enriched by the knowledge of how we have arrived at this point, but how exactly is the understanding that is provided by a diachronic perspective connected to the synchronic (and, in principle, independently presentable) senses of those ideals and values (cf. Gallie 1956b, 196–97)? Let us take another look at essentially contested ART.

When one claims or rejects a claim that something is "art," one is inevitably using the term in a contestable way because what one says can easily be recognized as appreciation or criticism from any of the historically manifested and (excessively) one-sided points of view (Gallie 1956a, 113–14). Gallie ends up claiming that this is brought about by the very nature of the arts as activities that are "ever expanding, ever reviving and advancing values inherited from a long and complex tradition" (ibid., 114). More generally,

> In any field of activity in which achievements are prized because they renew or advance a highly complex tradition, the point of view from which our appraisals are made—our concept of the achievement in question—would seem always to be of the kind I have called 'essentially contested'. Gallie 1956a, 114

Notwithstanding Gallie's curiously reverse formulation, the phenomenon that he is arguably describing is relatively straightforward and commonsensical: we humans engage in many activities or practices from which traditions of thought spring, traditions that are concerned with the best way to sustain and develop ideals or values which the selfsame activities and practices are perceived to manifest. When the aspects or features of an ideal or value make up a complex, or are perceived as such, the ideal or value admits various descriptions of what is of the utmost importance regarding it. Different descriptions espousing differing evaluations may result in the tradition itself becoming complex or branched. Gallie as-

sumes that the conceptualization of relevant ideals and values is at least partly mediated by traditions of thought: we learn to view things in a particular way from various cultural and historical sources, or as part of our everyday interactions with others who are similarly situated, and complex or branched traditions present us with multiple and often mutually exclusive options. This means that the concepts of our ideals and values are also complexly shaped by the past history or "the whole gamut of conditions" (Gallie 1956b, 196) that informs and guides us to endorse and conceptualize those ideals and values. What may at first sight look like an unfortunate confusion from a synchronic perspective may turn out to be an integral part of the social and intellectual fabric locally or universally. A diachronic or historical perspective is now required to separate the wheat from the chaff: we want to understand, indeed, we need to understand, which apparent confusions involve a continuing contestation that is of such significance to us that even our concept of the ideal or value reflects and represents that conflict.

Whether there really is, at the center of a human activity or practice, a singularly identifiable ideal or value that is collectively sustained and developed—or in different terms: it is normatively binding—is somewhat beside the point. What matters is that people appear to believe that certain exemplars and samples are embodying a deeper value. There may be no telling whether, in any given instance, it is really so. Psychological essentialism merely represents a belief that there are essences; whether one's knowledge about particular "essences" is accurate or not is a completely different matter (cf. Gelman and Wellman 1991, 229). In other words, "psychological essentialism refers not to how the world is but rather to how people approach the world" (Medin 1989, 1477). This means that the current theoretical framework for understanding ECCs cannot establish that having disputes that are centered around psychologically essentialized representations is necessarily a perfectly rational thing to do. The disputants perceive there to be an ideal or value that underlies each

concept use (or gives it a point[20]), and they disagree about what everyone should make of it.

The present understanding of ECCs also complements the way we understand both DCCs and NKCs. I have already mentioned the possibility that DCCs and NKCs may become essentially contested in suitable circumstances even if this might mean that such concepts should then be viewed as ECCs instead. In addition, we are getting a better sense of the workings of socially constructed DCCs especially. It is one thing to say that people associate a concept with a collection of features based on a value they perceive to be underlying the concept, but quite another to understand the process in which these features and the perceived underlying value come together as a basis for different categorizations that the concept licenses. For instance, think of ROCK MUSIC, which has been claimed to be a DCC (Knobe, Prasada, and Newman 2013). It is certainly not the case that we are free to associate rock music with any set of features or any value if we wish to employ the same concept with our fellows and thus share in their thought-processes. Instead, we have access to cultural information about rock music based on which we conceive of samples or exemplars as belonging to the same historical continuum that we perceive as embodying value that is characteristic to rock music. Sometimes concrete features (e.g., the sound that is centered on the amplified electric guitar; lyrics about social and political themes etc.) seem more relevant, sometimes a deeper value (e.g., rebelliousness[21]). Nevertheless, because DCCs have a structure similar to ECCs, there is reason to suspect that the kind of historical understanding that Gallie sought comes in handy also in the case of socially constructed DCCs.

---

[20] For different senses of "the point of a concept," see Queloz 2019. See also Pennanen 2021, sec. 18.2, for a discussion in the context of essential contestability.

[21] One way that DCCs may differ from ECCs is that they may perhaps be associated with *several* relatively distinguishable values that underlie a concept and tie concrete features together (e.g., authenticity and rebelliousness and perhaps more in the case of rock music). Whether this difference is real, or something that manifests in people's actual conceptual judgments, requires further empirical study.

I suspect that not everyone will agree with my current take on the nature of ECCs, so let me try to anticipate and briefly answer a couple of lines of criticism. First, one might want to object that the necessity of contestedness can be inferred from the empirical fact of contestedness only on pain of fallacy (cf. Ball 2001, 35) or that the modality of contestability is quite different from contestedness, and that I make a category mistake by appealing to empirical studies. However, just the same as a word- or term-usage is commonly taken as an indication of an underlying conceptual and/or cognitive structure, I do not see why systematic psychological studies that explicitly aim to reveal such structure(s) could not. A philosophical examination that adequately respects the rules of logic can continue from there, just as it would with any other information about the world. However, my case would be somewhat weakened if a (rational) philosophical intuition or insight about concept usage were markedly different or somehow more reliable than the layman's judgment—given that the three concept types look to share even more characteristics with each other, in practice, if NKCs also follow a dual character pattern. However, in absence of a convincing argument to the effect that a professional philosophical insight and the layman's judgment are different, one should minimally withhold from making that assumption (Machery 2017).

Evidence of what people's ordinary judgments regarding certain concepts are or in what ways they apply linguistic expressions that, for all we know, stand for these concepts, is very relevant in any case. Getting to the bottom of conceptual aspects of the intractable disputes of our time does not seem feasible without paying attention to the way people actually employ concepts, for example, to categorize items. From this perspective, disputing parties' conflicting judgments and their distinct patterns are something to be understood and explained, not explained away. Nevertheless, while there are established and relatively uncontroversial methods of testing people's conceptual judgments in psychology and cognitive science, none of the sort were utilized by Gallie nor do I employ them in this paper. The material question "Is there really that kind of concept?" is particularly hard one for a philosopher to answer positively, and often the only recourse is to

argue for the coherence and explanatory value of one's account. To get beyond a pure theory or stipulation requires more—not fewer—empirical studies that are well-thought and precise.

Second, some may find my invocation of essentialism objectionable. It is commonly presumed that Gallie wants to avoid a commitment to essentialism or that this is at least what he should do. According to one critical remark, Gallie "talks as if, behind each 'essentially contested concept', there was, hidden away in some Platonic heaven, a non-contested, unambiguously defined and fully determinate concept or exemplar" (Gellner 1974, 99). This type of metaphysics is commonly shunned today, and undoubtedly for good reason. So is it completely misguided to appeal to a form of psychological essentialism, let alone one dubbed as "Platonic essentialism?" Not at all, and there are other scholars too who have already come close to my position. For instance, Michael Stokes (2007) points out that requiring an exemplar enables a defense against the charge of Platonism, yet he wonders if it is possible to identify the important features of the exemplar without some intuitive understanding of an ideal type, in which case the exemplar would not offer a complete defense against such a charge (Stokes 2007, 690n22; compare with Gallie 1956a, 99–102). Stokes does not elaborate on specific forms that the intuitive understanding of the ideal type might assume; nevertheless, he holds that ECCs can be seen to admit different conceptions "because of continuing disputes about the most justifiable understanding of the values which underlie the concept" (Stokes 2007, 693).

The above points are, of course, very much in line with what I have been saying in this paper. The current framework significantly adds to the matter by (i) clarifying the structure of ECCs, (ii) illustrating by comparison that ECCs as a class of concepts is not as mysterious as might seem at first, and (iii) offering a way to track a conceptual mechanism that looks to be required by ECCs: psychologically essentialist categorization tendencies in everyday conceptual judgments need to be considered in conjunction with an externalist or historicist interpretation of essential contestability. There is no dubious metaphysics here; whatever it is that is "hidden" in an ECC—an ideal type, a value as a deep structure that gives point to a

category's features, or something similar—it is conceptualized into existence by concept-users themselves. For all the talk of Platonic essentialism, psychological essentialism and by extension the current theoretical framework are compatible with *not* accepting a type of Platonic idealism about our conceptual categories.

Finally, is there any reason to believe that my account of ECCs is consistent with any theory of concepts at all? Gallie himself was not satisfied with the prevalent method of seeking definitions in terms of necessary and jointly sufficient conditions for all concepts (Gallie 1956b, 185n3; see also Pennanen 2021, 32–34, 50, 98–100), or the notion that empirical sciences should provide the model for understanding concepts in other fields as well (Gallie 1956b, 168, 179, 197–98). ECCs exhibit features that may be viewed as more properly belonging either to the prototype theory, the exemplar theory, or the theory-theory, which all have challenged and to different extents replaced the classical theory of concepts.[22] The theory-theory connects especially well with psychological essentialism as it allows people to access a mentally represented theory when they make certain category decisions (Laurence and Margolis 1999, 46). It also coheres well with Gallie's choice to treat proposed theories and definitions as the concrete vehicles of essential contestability (e.g., Gallie 1956a, 112; 1964, 177; quoted in sec. 4). Psychological essentialism does not require a detailed understanding of the matter in question or clearly developed views about the nature of the property (Laurence and Margolis 1999, 46), and neither does the theory-theory. A theory behind an advocated concept use could also be a folk theory,[23] or perhaps mutually contesting concept-users just otherwise act as if their concepts contain "essence placeholders" (see Medin and Ortony 1989,

[22] For general features of these theories, see, e.g., Laurence and Margolis 1999 or Murphy 2002.

[23] Interestingly, Knobe, Prasada, and Newman speculate that conceptual representations of those employing DCCs may be shaped by "normative theories" about abstract values, theories which serve to unify certain category features rather than others (Knobe, Prasada, and Newman 2013, 255). For a tentative account of what this might mean in relation to Gallie's thesis, see Pennanen 2021, 371, n. 374, 434.

184). The latter option should be compatible with several other theories of concepts as well.

7. Conclusions

In this paper, I have examined whether ECCs have a dual character by comparing them to DCCs and NKCs. The answer is affirmative: there are striking similarities between ECCs and DCCs. First, categorizations made by employing ECCs and DCCs make use of two sets of criteria, descriptive and normative. It may be possible to cash the specific nature of these criteria in different ways, yet current findings support the conclusion that ECCs encode both a descriptive dimension as well as a somewhat independent normative dimension for categorization. Both ECCs and DCCs admit their users to dissociate the two dimensions in a way that licenses normatively guided categorizations "true X" and "not true X" in addition to more ordinary classifications "X" and "not X."

Some empirical studies on ordinary speakers' use of natural kind terms and related conceptual judgments suggest that conceptual judgments involving NKCs also evidence a dual character pattern. Unlike with DCCs and ECCs as they are originally presented by Gallie, assuming the presence of the two sets of criteria goes against an established theory in the case of NKCs, which gives one pause. Although there may be good reasons to stick to a textbook definition with NKCs or natural kind terms, for the present purposes it is enough to identify a common pattern in ordinary judgments between these concept types as it renders ECCs less mysterious as a class of concepts. I also discussed the possibility that the terms denoting ECCs and NKCs respectively are species of a single semantic genus. While natural kind terms are correctly applied to something if and only if it bears the relation of "belonging to the same kind as" to samples or exemplars that have played a historical role in the use of the term, in the case of ECCs the operative relation is "being the heir of" a component of which is the relation "being part of the same tradition." This goes a long way towards explaining why historical understanding is required in the case of ECCs: determining the matter requires assessing whether conflicting

uses of concepts are traceable and faithful to past exemplars and samples and, as such, whether they are intelligible.

The contestability that arises from conflicting verdicts on the applicability of a concept based on the two sets of criteria is not really discussed in the literature on ECCs, literature that mostly focuses, in Gallie's footsteps, on contestation that plays out on the normative side. Still, disagreements over which set of criteria should be used for categorization in a given case is still a live possibility in disputes involving ECCs given their dual character. Then again, studies on DCCs have hitherto overlooked the possibility that a dispute could arise over how to understand a concept's underlying abstract value. Concepts such as ART, SCIENCE/SCIENTIST, and CHRISTIANITY /CHRISTIAN that have been independently put forward as candidates for being DCCs are also examples of ECCs that Gallie mentions. However, the assumption that the values underlying concepts can be contested does not come without a cost: identifying contestability at the level of a concept's structure introduces the unity problem—i.e., disputing parties may not be employing/contesting the same concept—which is difficult to solve, and this potentially applies to both DCCs and ECCs. Essential contestability appears to constantly challenge the acceptable boundaries of conceptual identity and variation, but it may also lead one to question whether the insight behind Gallie's thesis can even be captured by the view that understands concepts *qua* concepts as the origin of essential contestability. Therefore, and somewhat paradoxically, I cannot give a conclusive answer to the question of whether DCCs could become essentially contested given that the very notion of such contestedness/contestability is somewhat questionable. Nevertheless, assuming that the unity problem is solvable or that it can be worked around, it may still be separately advisable to classify "DCCs" that become essentially contested more simply as ECCs. The final determination of what is what depends heavily on one's background view or theory of concepts.

Second, I have proposed that ECCs are accompanied with a form of psychological essentialism, dubbed "Platonic Essentialism." Gallie's commitment to essentialism has been critically suggested in the literature before, yet after comparing the features of NKCs and DCCs, and then considering ECCs

together with DCCs, it becomes possible to see ECCs in a different light. Now, an ECC is involved in a dispute when mutually contested and contesting uses of a concept are faithful to exemplars and samples, all of which belong or are claimed to belong to the same tradition (or one of its branches) on the grounds that they embody and/or manifest the same abstract value or normative ideal. Contesting uses both aim and are claimed to be true representations of a historically embodied normative ideal or value, which exhibits psychological essentialism. I defended this view against the charge of taking concepts to be immutable and eternal entities, and I also gave a brief answer to the objection that the necessity of contestation, or a concept's contestability, cannot be grounded in empirical facts about concept employment.

The current account of ECCs is able to take seriously the criticism that an advocate of ECCs might end up subscribing to Platonic idealism while incorporating essentialism in a modified psychological form as a key factor in the overall explanation. This also means that ECCs are value-laden not necessarily because that which falls under a concept's extension is intimately connected to a value, or that the value somehow inheres in the concept, but because concept-users simply consider certain exemplars and samples of the concept as manifestations or realizations of the ideal or value. People's normative differences are then reproduced in the ways they apply the concept. This perspective of essential contestability is only concerned with the way disputing parties conceptualize the contested issue in question, and contestability thus becomes a matter that originates in their beliefs, attitudes, and practices. The current theoretical framework is potentially compatible with multiple theories of concepts, although it leans towards the theory-theory view or some hybrid-view that entails it.

The conceptual architectures of ECCs, DCCs, and NKCs are similar enough to suspect that DCCs and even NKCs could also become involved in contestation that is much like Gallie describes in the case of ECCs. It arguably requires the right conditions, though, and some of the conditions that should be different for there to be essential contestability may be integral to employing the type of concept in question. Changes in a concept's architecture may therefore mark shifts

from one concept type to another. Establishing these effects requires further study, both theoretical and empirical. Given that essential contestability is a phenomenon that is intimately tied to both culture and history, separating contributing factors from everything else that is or could be going on is not an easy task. Recognizing the dual character of ECCs is a start.

Reframing essential contestability in terms of psychological essentialism is a fresh perspective to the phenomenon of essential contestability which also points toward an improved, full essential contestability thesis. The new framework is compatible with most of the insights of Gallie's original thesis while steering clear of some of its logical problems. By grounding the structure of ECCs in certain conceptual operations of disputants rather than in the joints of reality, my account more generally suggests that the dispositions of the parties to a dispute are crucial for understanding essential contestability.

University of Jyväskylä

References

Ball, Terence (2001). "From Hobbes to Oppenheim: Conceptual Reconstruction as Political Engagement." In Ian Carter and Mario Ricciardi (eds.), *Freedom, Power and Political Morality: Essays for Felix Oppenheim*. London: Palgrave Macmillan UK, 20–38.

Besson, Samantha (2005). *The Morality of Conflict: Reasonable Disagreement and the Law*. Oxford: Hart Publishing.

Boromisza-Habashi, David (2010). "How Are Political Concepts 'Essentially' Contested?" *Language & Communication* 30(4), 276–84.

Braun, David, and Theodore Sider (2007). "Vague, So Untrue." *Noûs* 41(2), 133–56.

Bryant, Christopher G. A. (1992). "Conceptual Variation and Conceptual Relativism in the Social Sciences." In Diederick Raven, Lieteke van Vucht Tijssen, and Jan de Wolf (eds.), *Cognitive Relativism and Social Science*. New York: Routledge, 51–67.

Burg, Wibren van der (2017). "Law as a Second-Order Essentially Contested Concept." *Jurisprudence* 8(2), 230–56.

Collier, David, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu (2006). "Essentially Contested Concepts: Debates and Applications." *Journal of Political Ideologies* 11(3), 211–46.

Connolly, William E. (1993). *The Terms of Political Discourse*. Third edition. Oxford: Blackwell.

Criley, Mark Edward (2007). "Contested Concepts and Competing Conceptions." Doctoral dissertation, University of Pittsburgh.

Del Pinal, Guillermo, and Kevin Reuter (2015). "'Jack Is a True Scientist': On the Content of Dual Character Concepts." In *Mind, Technology, and Society*, Austin, 23 July 2015–25 July 2015, 554–559.

Del Pinal, Guillermo, and Kevin Reuter (2017). "Dual Character Concepts in Social Cognition: Commitments and the Normative Dimension of Conceptual Representation." *Cognitive Science* 41(S3), 477–501.

Evnine, Simon J. (2014). "Essentially Contested Concepts and Semantic Externalism." *Journal of the Philosophy of History* 8(1), 118–40.

Freeden, Michael (1996). *Ideologies and Political Theory: A Conceptual Approach*. Oxford; New York: Clarendon Press: Oxford University Press.

Freeden, Michael (2004). "Editorial: Essential Contestability and Effective Contestability." *Journal of Political Ideologies* 9(1), 3–11.

Gallie, W. B. (1956a). "Art as an Essentially Contested Concept." *The Philosophical Quarterly (1950)* 6(23), 97–114.

Gallie, W. B. (1956b). "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56, 167–98.

Gallie, W. B. (1964). *Philosophy and the Historical Understanding*. London: Chatto & Windus.

Garver, Newton (1987). "Violence and Social Order." In Ota Weinberger, Peter Koller, and Albert Schramm (eds.), *Philosophy of Law, Politics, and Society. Proceedings of the 12th International Wittgenstein Symposium*. Vienna: Hölder-Pichler-Tempsky, 218–23.

Gellner, Ernest (1974). "The Concept of a Story." In Ernest Gellner, *Contemporary Thought and Politics*. London and Boston: Routledge & Kegan Paul, 95–112.

Gelman, Susan A., and Henry M. Wellman (1991). "Insides and Essences: Early Understandings of the Non-Obvious." *Cognition* 38(3), 213–44.

Grafstein, Robert (1988). "A Realist Foundation for Essentially Contested Political Concepts." *The Western Political Quarterly* 41(1), 9–28.

Gray, John (1978). "On Liberty, Liberalism and Essential Contestability." *British Journal of Political Science* 8(4), 385–402.

Gray, John (1983). "Political Power, Social Theory, and Essential Contestability." In David Miller and Larry Siedentop (eds.), *The Nature of Political Theory*. Oxford & New York: Clarendon Press, 75–101.

Hampton, James A. (1979). "Polymorphous Concepts in Semantic Memory." *Journal of Verbal Learning and Verbal Behavior* 18(4), 441–61.

Haukioja, Jussi, Mons Nyquist, and Jussi Jylkkä (2021). "Reports from Twin Earth: Both Deep Structure and Appearance Determine the Reference of Natural Kind Terms." *Mind & Language* 36(3), 377–403.

Ingram, Peter (1985). "Open Concepts and Contested Concepts." *Philosophia* 15 (1–2), 41–59.

Knobe, Joshua, Sandeep Prasada, and George E. Newman (2013). "Dual Character Concepts and the Normative Dimension of Conceptual Representation." *Cognition* 127(2), 242–57.

Lakoff, George (1973). "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts." *Journal of Philosophical Logic* 2(4), 458–508.

Laurence, Stephen, and Eric Margolis (1999). "Concepts and Cognitive Science." In Eric Margolis and Stephen Laurence (eds.), *Concepts: Core Readings*. Cambridge, MA: MIT Press, 3–81.

Leslie, Sarah-Jane (2015). "'Hillary Clinton Is the Only Man in the Obama Administration': Dual Character Concepts, Generics, and Gender." *Analytic Philosophy* 56(2), 111–41.

Liao, Shen-yi, Aaron Meskin, and Joshua Knobe (2020). "Dual Character Art Concepts." *Pacific Philosophical Quarterly* 101(1), 102–28.

Lukes, Steven (1977). "A Reply to K. I. Macdonald." *British Journal of Political Science* 7(3), 418–19.

Lukes, Steven (2005). *Power: A Radical View*. Second edition. Basingstoke: Palgrave Macmillan.

Machery, Edouard. (2017). *Philosophy within Its Proper Bounds*. New York, NY: Oxford University Press.

Medin, Douglas L. (1989). "Concepts and Conceptual Structure." *American Psychologist* 44, 1469–81.

Medin, Douglas L., and Andrew Ortony (1989). "Psychological Essentialism." In Stella Vosniadou and Andrew Ortony (eds.), *Similarity and Analogical Reasoning*. New York, NY, US: Cambridge University Press, 179–95.

Murphy, Gregory L. (2002). *The Big Book of Concepts*. Cambridge, Mass.: MIT Press.

Newey, Glen (2001). "Philosophy, Politics and Contestability." *Journal of Political Ideologies* 6(3), 245–61.

Newman, George E., and Joshua Knobe (2019). "The Essence of Essentialism." *Mind & Language* 34(5), 585–605.

Pennanen, Joonas (2021). "Essentially Contested Concepts: Gallie's Thesis and Its Aftermath." JYU dissertations, University of Jyväskylä.

Putnam, Hillary (1975). "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7, 215–71.

Reuter, Kevin (2019). "Dual Character Concepts." *Philosophy Compass* 14(1), e12557.

Ricciardi, Mario 2001). "Essential Contestability and the Claims of Analysis." In Ian Carter and Mario Ricciardi (eds.), *Freedom, Power and Political Morality: Essays for Felix Oppenheim*. London: Palgrave Macmillan UK, 39–56.

Rosch, Eleanor, and Carolyn B Mervis (1975). "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7(4), 573–605.

Ruben, David-Hillel (2010). "'W.B. Gallie and Essentially Contested Concepts': Re-Reading of W.B. Gallie, 'Essentially Contested Concepts', *Proceedings of the Aristotelian Society* (1956) 167-198." *Philosophical Papers* 39(2), 257–70.

Ruben, David-Hillel (2013). "Traditions and True Successors." *Social Epistemology* 27(1), 32–46.

Stokes, Michael (2007). "Contested Concepts, General Terms and Constitutional Evolution." *Sydney Law Review* 29(4), 683–712.

Swanton, Christine (1985). "On the 'Essential Contestedness' of Political Concepts." *Ethics* 95(4), 811–27.

Tobia, Kevin P., George E. Newman, and Joshua Knobe (2020). "Water Is and Is Not H_2O." *Mind & Language* 35(2), 183–208.

Väyrynen, Pekka (2014). "Essential Contestability and Evaluation." *Australasian Journal of Philosophy* 92(3), 471–88.

Waldron, Jeremy (1994). "Vagueness in Law and Language: Some Philosophical Issues." *California Law Review* 82(3), 509–40.

Weitz, Morris (1972). "Open Concepts." *Revue Internationale de Philosophie* 26 (99/100 (1/2)), 86–110.

Wright, Crispin. 2003. *Saving the Differences: Essays on Themes from Truth and Objectivity*. Cambridge, Mass.; Harvard University Press.

Zimmerling, Ruth. 2005. *Influence and Power: Variations on a Messy Theme*. Dordrecht, Netherlands: Springer.

Unified (Enough) Metasemantics for Expressivists

TEEMU TOPPINEN & VILMA VENESMAA

We discuss a challenge for expressivism in metaethics. According to expressivism, the meaning of normative sentences is explained by their playing a *practical role*, or by facts about what desire-like, or action- or attitude-guiding states of mind, normative sentences express. We first explain how expressivism can be understood as a view about the *metasemantics* of normative language (section 1). The challenge, which we may call *the problem of diverse uses* (Väyrynen 2022), is based on the simple observation that while terms such as "good" or "ought" plausibly have a unified meaning across a wide variety of different uses, not all uses of sentences that contain these terms seem to play a suitably practical role. How, then, can the expressivist explain the meaning of such sentences by appealing to the idea that they play a practical role (section 2)? We suggest that expressivists can deal with this challenge. Our response is based on two ideas. First, understanding expressivism as a view in metasemantics rather than in semantics creates space for the possibility that both the practical and the descriptive uses of normative terms might carry the same meaning. This requires adopting a metasemantics that has some complexity, which leads to what we may call *the problem of disunified metasemantics* (Wodak 2017, Väyrynen 2022). However, we argue that this problem may nevertheless be dealt with, given that the extra complexity is required in order to capture the relevant phenomena (section 3). Second, in order to avoid a remaining challenge that we may call *the problem of unexplained metasemantic coincidence* (Wodak 2017), the expressivist account should take a certain kind of form. We suggest that a

view called *relational expressivism* holds promise with regard to the prospect of giving a unified enough metasemantics for normative language that doesn't rely on unexplained coincidences (section 4). Finally, we briefly conclude (section 5).

1. Expressivism and the practical role of normative language

By the term "normative language" we may single out, roughly, those chunks of language that centrally deploy terms such as "good," "ought," and "reason," or whatever it is that these terms translate to in languages other than English. Examples of normative sentences would then be sentences such as "Knowledge is good in itself," "We ought to ban Nazi symbols," or "There's some reason to eat cars."

Normative terms, or terms such as "good," "ought," and "reason," play a *practical role* in our thought and talk. A term plays a practical role, we might say, when its use normally expresses the speaker's practical attitudes of some appropriate kind. But the meaning of "expresses," and consequently, the idea of a practical role, can be understood in very different ways. One possible view would be that normative language expresses practical attitudes as a *broadly semantic* matter, or in virtue of suitable linguistic conventions. According to an alternative view, uses of normative language express such attitudes as a matter of *pragmatics*.

As an example of a view of the latter sort, it might be suggested that normative claims carry a *generalized conversational implicature* to the effect that the speaker has certain attitudes. When some claim conveys, via a generalized conversational implicature, that the speaker has a certain attitude, the suggestion that the speaker has this attitude is not a part of what is said, and can be "canceled," but can nevertheless be assumed to be true in the absence of a special context. An example might be provided by the sentence "They drank some of the tea," which, in the absence of special circumstances, implicates, but doesn't say, that the person picked out by "they" did not drink *all* of the tea. (As noted, what is thereby pragmatically conveyed by the use of the sentence is, in this sort of case, cancellable in that it would not be linguistically inappropriate to say, "They drank some of the tea—indeed,

they drank it to the last drop." On appeals to generalized conversational implicatures in relevant contexts, see Strandberg 2011, Fletcher 2014.)

Expressivism, by contrast, is an example of a view that construes the practicality of normative language as a broadly semantic matter. According to expressivism, the meaning of normative sentences is explained by their playing a practical role, or by facts about what desire-like or action- or attitude-guiding states of mind normative sentences express (e.g., Blackburn 1998, Gibbard 2003, Ridge 2014). This is plausibly understood as a thesis about *metasemantics*. That is, expressivism plausibly offers an explanation for why normative terms and sentences have the meaning that they do have, or an account of what it is in virtue of which they have their meaning.

Here the notion of expression is different from the pragmatic one mentioned above and illustrated with reference to generalized conversational implicatures. The expression relation, as understood here, in the context of expressivism, is not a matter of communicating a piece of information. Rather, it figures in the explanation of how normative language gets to have its meaning. What is it, then, for a sentence to express, in the relevant sense, some state of mind or attitude? Our project, here, is not that of developing an account of the expression relation. But perhaps something along the lines of the account proposed by Michael Ridge (2014, 109) will do for illustrative purposes:

> A declarative sentence '*p*' in sense *S* in a natural language *N* used with assertive force in a context of utterance *C* expresses a state of mind *M* if and only if conventions which partially constitute *N* dictate that someone who says '*p*' in sense *S* in *C* with assertive force is thereby *liable* for being in state *M*.

This proposal requires some clarification. What is it to use a sentence with assertive force? Presumably, the idea is that using a sentence with assertive force just is a matter of using the sentence in such a way that one is liable or can be held accountable—in the light of the relevant linguistic conventions—for being in *some* state *M*, where this state must be of a certain type: a *belief-like* state or commitment, or a state or

commitment that has a *telos*, or function, of matching the world.[1]

Now, one might worry that this is not an expressivist-friendly proposal. For according to expressivism, normative sentences express practical states of mind or commitments, where these are to be somehow *contrasted* with representational beliefs, the *telos* or function of which is to represent the world as being this or that way. How, according to expressivism, could these sentences then be used with assertive force, if this is a matter of using a sentence in such a way that one is liable for being in a belief-like state, or for having a belief-like commitment? This kind of worry is easily disposed of. Expressivists should agree that the relevant states of mind or commitments are belief-like, in the relevant sense, despite not being representational beliefs the *telos* or function of which is to represent the world as being this or that way, normatively speaking. The relevant distinction can be drawn in many ways, but we do not wish to commit ourselves to any particular way of doing so, or to discuss this issue in any more detail here.[2] It suffices, for the present purposes, that expressivists may propose to understand the states or commitments expressed by normative sentences as being belief-like in some sense such that their expression amounts to an assertive use, and yet reject the view that these states or commitments could be understood simply as being in the business of representing the "normative bits of reality" (as Sinclair 2021 puts it).

Clearly more could and should be said by someone, somewhere, on what expressivists could and should say about the expression relation. But perhaps this suffices here (for more on the topic, see, e.g., Schroeder 2008, Sinclair 2021,

---

[1] Why "in sense *S*"? Presumably this is meant to restrict the set of assertive uses of the sentence, at issue, to those in which the sentence is used assertively in a certain type of way. So, one sentence, even when used assertively, might express one kind of state of mind in one type of use, and another kind of state of mind in another type of use. For example, "Allowing the use of Nazi symbols is wrong," when used assertively and in a practical role, might express one kind of state of mind. And the same sentence might express a different kind of state when used, in a different kind of role, to report the mores of the surrounding society.

[2] For discussion, see, e.g., Sinclair 2007, Schroeder 2010, Ridge 2014.

Ch. 3). This rough account allows us to see how expressivism neatly captures the practicality of normative language. Normative language plays a practical role because that's what it's for. Normative terms mean what they mean because of their contribution to sentences that express practical attitudes, states, or commitments.

We have suggested that expressivism is best understood as a thesis in metasemantics, or as the thesis that the meaning of normative sentences is *explained* by facts about what states of mind these sentences express, in the relevant sense. By contrast, we may think of *semantics*, roughly, as giving a systematic account of what some set of sentences, *S*, mean, where this account is given in terms that we already understand. So, if we wanted to give a semantics for a snippet of the normative chunk of English, we could try saying, for instance, things such as the following:[3]

> The meaning of "good" is a function from objects that gives the value *true* just in case the contextually relevant object has the property of goodness.
>
> The meaning of "ought" is a function from agents and actions that gives the semantic value *true* just in case the contextually relevant agent has an obligation to perform the contextually relevant action.

A sentence such as "Knowledge is good" would, then, be true, just in case the thing picked out by "knowledge"—knowledge—has the property of goodness. And a sentence such as "Tove ought to turn down the volume" would be true just in case the person referred to by "Tove" would have an obligation to turn down the volume in the relevant context. Or perhaps we would rather wish to say something along the following lines:

> The meaning of "good" is a function from objects that gives the value *true* just in case the contextually relevant object is highly

[3] Our formulations of the semantic theses below have drawn inspiration from the way some relevantly similar toy accounts are formulated in Chrisman 2016.

> ranked by certain standards, where the relevant standards are fixed by the context.[4]
>
> The meaning of "ought" is a function from a proposition that gives the semantic value *true* just in case the proposition is true in all of the worlds that are ranked highest by standards of a certain sort, among a set of worlds restricted in certain ways, where the relevant standards and restrictions are contextually determined.[5]

A sentence such as "Knowledge is good" would, then, be true, just in case the thing picked out by "knowledge"—knowledge—would be highly ranked by certain contextually determined standards. And a sentence such as "Tove ought to turn down the volume" would be true just in case Tove would turn down the volume in all of the worlds consistent with certain contextually determined background conditions and highly ranked by certain contextually determined standards.

To supplement our semantics for normative language, we could then add rules that allow us to use certain terms to create more complex sentences that have two or more simpler sentences as their parts. So, "and," for instance, would contribute a function from propositions that gives the semantic value *true* just in case all of the relevant propositions are true. And so on.

Of course, this only gives us two candidate accounts of semantics for the tiniest of snippets of the normative chunk of English. Furthermore, these two accounts undoubtedly are poor candidates, too. But that's OK. They are just toy models. The point, here, is simply that expressivists could sensibly think that *some* such truth-conditional semantics for the normative chunk of English can be given. Thus far, in providing a semantics for normative language, we wouldn't have needed to invoke, for instance, anything like the idea that the ac-

---

[4] This is inspired by the account of the meaning of "good" in Ridge 2014, 26.

[5] This is a Kratzerian account of the meaning of "ought" (see Kratzer 2012), the formulation of which draws from Chrisman 2016, 86. For discussion of the semantics of "ought," including Kratzer's view, see also, e.g., Bronfman & Dowel 2018, Carr 2018.

ceptance of a normative sentence is a matter of representing the world as being a certain way, rather than a matter of having some desire-like attitude. We wouldn't have needed to appeal to any specific metaphysical account of the normative properties. The connection between semantics, in the sense outlined above, and many of the issues debated in metaethics doesn't seem to be especially close or direct (for a more detailed development of this point, see, e.g., Ridge 2014, Chrisman 2016, Sinclair 2021).[6]

We now are in a position to say some more about what expressivism might be taken to amount to. According to expressivism, again, the meaning of normative sentences is explained by their expressing certain suitably practical states of mind or commitments. That is, their meaning is explained by the fact that, when used assertively, the speaker of a sentence of the relevant kind can be held accountable, by linguistic convention, for being in some appropriate belief-like, yet practical, state of mind. On this kind of view, then, the sentence "Allowing the use of Nazi symbols is wrong" might express (appropriately belief-like) opposition to the use of Nazi symbols. This would be so in virtue of the fact that the relevant linguistic conventions would dictate that, when used assertively, the speaker of the sentence could be held accountable for being opposed (in an appropriately belief-like manner) to allowing the use of Nazi symbols.

2. The Problem of Diverse Uses

The appeal to the practical role of normative language in explaining its meaning gives rise to a challenge which we may, following Pekka Väyrynen (2022), call *the problem of diverse uses*. The challenge is based on the simple observation that not all uses of sentences that contain these terms seem to play

[6] The point is merely that expressivism seems compatible with such semantic views. This is not at all to suggest that expressivists thereby escape having to deal with various metaphysical issues. For instance, given that there really are normative properties, it seems fair to ask the expressivist what such properties are like (whether they are, for example, *sui generis*, or reducible to properties that may also be ascribed by descriptive judgments; see, e.g., Bex-Priestley forthcoming).

a suitably practical role. Väyrynen (2022) characterizes the problem as follows:

> If the practical role of these terms were a part of their conventional profile in a language, it should not be subject to […] exceptions but instead should be present in all literal uses in normal contexts. This raises what I will call the Problem of Diverse Uses: How do you reconcile the diversity of uses to which […] normative terms may be put with the claim that their association with their normative roles is broadly semantic? The problem prompts a challenge: either offer some plausible explanation of cases where the relevant practical upshots are absent that reconciles these claims, or else do not build such upshots into our overall semantic theory for […] normative terms (182–183).

We should look at some examples of candidates for non-practical uses of "ought." Väyrynen (2022, 182–183) offers a useful selection. Consider, then, the following:

> (1) One ought to prioritize profit over fairness. But is that really the thing to do?

In (1), the ought-claim may, in a suitable context, be rightly understood as a claim about what follows from capitalist values or standards. It might be clear from the preceding discussion, or from the pins on the speaker's jacket (Väyrynen 2022, 192), that such standards are not the speaker's standards. Plausibly, the ought-claim in (1) need not, then, play any practical role for the speaker. Or consider:

> (2) Client: What is my legal obligation, and what do you expect me to do?
>
> Lawyer: You have to report your liability, but I do not know if you will; you may prefer to push the limits of the law and just conceal it.

Here the client and the lawyer discuss legal oughts and musts, but they might not end up giving such oughts and musts any weight in their practical deliberation. Väyrynen (2022, 182–183) provides more examples:

> (3) It would be wrong to kill. But I'm ok with killing and do not feel bad about it.

(4) I ought to finish grading. I have absolutely no intention to do so, though.

(5) I should do the shopping today (as far as I know).

In all these cases, we may, given a suitable context, understand the relevant normative terms ("wrong," "ought," "should") in relation to standards that may not engage the speaker, or have any motivational significance for them.

Now, one thing that the expressivist could say is, of course, that "ought" means different things in different contexts. Sometimes it has a descriptive meaning. In these descriptive uses, ought-claims just report how things are ranked according to certain descriptively specifiable standards. Such claims need not play any practical role. In other contexts, though, "ought" has a very different kind of, practically charged, meaning. Expressivists have indeed sometimes suggested just this. A. J. Ayer (1936, 105–106), for instance, distinguishes between the "normative ethical symbols" and the "descriptive ethical symbols," which are "commonly constituted by signs of the same sensible form," but make a very different contribution to the meanings of sentences. However, this is not a promising route for the expressivist. It is a striking feature of normative terms such as "good," "ought," and "reason," that they *all* have both practical and non-practical uses, and that the patterns of their use are very similar across a range of languages. It is incredible that this would be due to normative terms being simply ambiguous. (Mackie 1977, 51, Chrisman 2016, Ch. 2.3; for warnings regarding a reckless postulation of lexical ambiguities, see also, e.g., Thomson 2008, Finlay 2014, Wodak 2017).

We do not wish to reject the possibility that normative words would turn out to be ambiguous in their meaning, possibly in a variety of interesting ways. But it does seem that the working hypothesis and the default position should be that there is significant unity to the meaning of the various uses of normative terms. In particular, and importantly in the present context, normative terms such as "ought" plausibly have a unified meaning across both practical and non-practical uses.

The desirability of a rather unified account of the meaning of normative terms provides one important reason for why it

makes sense to understand expressivism as a view in metasemantics.[7] For a metasemantic construal of expressivism creates space for a response to the problem of diverse uses. If expressivism is understood as a view in metasemantics, space opens for the following possibility: perhaps, in some cases, normative terms mean what they mean because of their contribution to sentences that express a practical state of mind; perhaps, in some other cases, normative terms have this same meaning because of their contribution to sentences that express another kind of, non-practical, state of mind.

For example, in the previous section, we mentioned the following toy semantics for "ought":

> The meaning of "ought" is a function from a proposition that gives the semantic value *true* just in case the proposition is true in all of the worlds that are ranked highest by standards of a certain sort, among a set of worlds restricted in certain ways, where the relevant standards and restrictions are contextually determined.

If this kind of semantics for "ought" is correct, then, when someone says, for example, that we ought to ban the use of Nazi symbols, their statement means that the use of Nazi symbols is banned in all of the worlds compatible with two contextual restrictions: First, the *modal base* restricts the set of worlds we are considering to those worlds that are compatible with whatever background conditions are determined by context *c* (for instance, those worlds in which it is possible for us to use Nazi symbols). Second, the *ordering source* further restricts the relevant set of worlds to those worlds which are ranked as best by some ordering over worlds, in accordance with whatever standards are determined by *c* (for instance, those worlds in the modal base that are best according to the correct standards of practical reason, or to give another ex-

[7] It's not the only one. We have already noted that going metasemantic seems to allow expressivists to adopt non-revisionary, standard views in semantics, which is nice also for other reasons. There are also reasons for understanding expressivism as a metasemantic view that are not narrowly semantic. One such reason, having to do with our knowledge of normative supervenience, is given in Venesmaa 2021.

ample, those worlds that are best according to the moral standards accepted in the speaker's community).

Now, this is all very toy-ish. But that's fine. What is important here is that expressivists can adopt some such semantic story with regard to "ought" (and the rest of normative language). Or better: it's not instantly obvious that they cannot do so. For they can suggest that while statements about what ought to be done, or about what ought to be, more generally, always have a meaning of this sort—a meaning captured by something like the toy account above—their expressivist account of normative language is entirely compatible with this. They may propose that expressivism helps to explain why the "ought"-sentences have the kind of meaning that they do have. More precisely, they may propose that the expressivist account explains whatever meaning "ought"-sentences have, according to the unified semantics, in some contexts, but not in others. This is why some of the "ought"-sentences are practical while others aren't. Or that's what metasemantic expressivism allows one to say.

This, by itself, is a very abstract point about the kind of structure that an expressivist proposal might take. It is one thing to point out that this kind of structure is, in principle, available to be utilized. It is another thing, entirely, to outline an expressivist view that has this kind of structure and that would be attractive. In the next two sections, we first present two challenges for the kind of expressivist response to the problem of diverse uses, according to which normative terms have a unified meaning, but this meaning is given an expressivist explanation only when the relevant terms are used in a practical role. We then articulate an expressivist view that is well-positioned to exploit the availability of this kind of response.

3. The problem of disunified metasemantics

The idea that we might be able to combine a fairly unified semantics for normative terms—one that captures a wide range of both practical and non-practical uses—with an expressivist metasemantics that only explains the meaning of normative terms in some of their uses is not a new one. The space for this kind of move has been explored before (see, in

particular, Ridge 2014). But the idea has met with some skepticism. We next wish to address the interesting articulations of such skepticism by Pekka Väyrynen (2022) and Daniel Wodak (2017).

Väyrynen argues that the existing expressivist metasemantics for normative language "do not support the claim that the practical role of such language is a distinctive and particularly significant feature of its meaning" (2022, 200). He reaches this conclusion through considering the different ways in which the practical role of normative language might figure in its metasemantics. Let us suppose, then—along with Väyrynen—that "ought" has a Kratzerian semantics such as the one that we have used as our toy semantics above (cf. Väyrynen 2022, 189–190). There are two options, Väyrynen suggests, with regard to understanding the nature of the metasemantic work that is done by the idea that normative sentences (sometimes) play a practical role. The first one "has to do with the metasemantics of the context-sensitivity of *ought*":

> Perhaps its practical role contributes to explaining its semantic value specifically in its committal uses. […] Whether a use is committal or not is a difference in context. We might then think that when *ought* is used in a committal way, this can make a difference to the values of its contextual parameters [the modal base and the ordering source]. In this way, the practical role of *ought* might contribute to explaining its semantic value in some cases but not others. (Väyrynen 2022, 190–191)

This is, indeed, one kind of metasemantic work done by the practical role of normative language. Whether an "ought"-sentence plays this kind of role contributes to determining its semantic value in a context. However, we agree with Väyrynen that this isn't all the metasemantic work that expressivists should take to be done by the idea that normative language sometimes plays a practical role. The expressivist idea is not merely that the practical role of normative language sometimes helps to determine the values of the contextual parameters of an "ought." This much could be agreed upon, for instance, by someone, according to whom the meaning of "ought"-sentences is always to be explained by their being expressive of robustly representational beliefs

about how various scenarios compare with regard to some contextually specified standards. This kind of representationalist, non-expressivist metasemantics could be combined with the idea that the relevant standards are sometimes determined by "ought" playing a practical role for the speaker in the context. The occasionally practical role of "ought" could then, in this way, contribute to determining the ordering source, and the semantic value, for some uses of "ought."

However, this kind of work in the explanation of meaning would not be everything that an expressivist needs from the practical role of normative language. Rather, according to a metasemantic expressivist, "ought" (for example) has the kind of semantics that it has (a Kratzerian semantics, say) because it plays a practical role. This would seem to correspond to Väyrynen's second proposal with regard to how the metasemantic significance of the idea that normative language plays a practical role could be understood. On this proposal, as Väyrynen puts it, the "practical role of *ought* is part of what explains why the dominant sort of formal models for modal language provide a good descriptive semantics for terms like *ought* in the first place."

It is important to emphasize, though, that this proposal should not be understood as suggesting that the meaning of "ought" is always, in every context, explained by its serving a practical role. We have granted, in the previous section, that the uses of "ought" are diverse, and not always practical. So, the expressivist proposal should be that the practical role of "ought" is part of what explains why normative language has the kind of semantics that it has *in those cases in which it does play a practical role*. The thought would be, then, that the meaning of "ought" remains constant across both practical and non-practical uses, but is only explained (in part) by the practical role of "ought" in the uses of the former sort. In both sorts of uses, the given semantics would be a good model for "ought" "because it appropriately mirrors the structure of mental states that *ought* expresses" (Väyrynen 2022, 193). In the non-practical uses, the meaning of "ought" would be explained by the fact that the relevant sentences express certain representational beliefs or commitments; in the practical uses, it would be explained by the fact that the relevant sentences express practical commitments or states of mind.

Väyrynen is not happy with this kind of suggestion. He writes:

> Explaining noncommittal uses only requires invoking theoretical commitments and cognitive states. By parity, that should suffice also for explaining committal uses. The standard semantics does not care about this distinction between uses. So, on the face of it, explaining why it is a good model for *ought* should not require invoking practical role. (It really is dialectically significant if committal and noncommittal uses of ought are uniform in their descriptive semantics!) If that is right, it would complete my case that nothing in the standard semantics for *ought* supports treating those uses that are associated with a practical role as semantically or metasemantically exceptional. (Väyrynen 2022, 193)

We have granted that in noncommittal or non-practical uses, the meaning of "ought" can be explained without appealing to the idea of a practical role. Väyrynen suggests that since it must then be possible to explain the basic semantic structure of "ought" without invoking practical role, and since we must, in any case, do so in the case of non-commital uses, we should also do so in the case of committal or practical uses. Why? Presumably, because this is the option favored by considerations of simplicity and uniformity; given that two metasemantic views do an equally good job in explaining why the standard Kratzerian view provides (what we are assuming is) a good model for "ought," but one is more simple and unified in that it doesn't invoke different kind of mental states in explaining its practical and non-practical uses, then this is a point in its favor.

As we understand Väyrynen's objection, the same objection is raised also by Wodak (2017) who targets Ridge's (2014) attempt to formulate an expressivist metasemantic theory that would vindicate unified semantics for "ought" and other normative terms:

> First, Ridge must concede that there is a viable non-expressivist explanation of why "ought" means *Z* in a wide variety of uses. This explanation is non-expressivist insofar as it appeals to robustly representational beliefs. And it is viable in that it explains: why "ought" means *Z*; how context selects the relevant

> ordering source; how competent speakers use "ought" to communicate, coordinate, and collect information; and how speakers disagree even in the face of systematic differences in their criteria for applying words (like "legally ought"). Once that viable non-expressivist explanation is on the table for some uses, why not offer it across the board? A unified meta-semantics is preferable, if only for the sake of parsimony. (Wodak 2017, 284)

However, what we want is not just the most simple and uniform metasemantic theory that explains why the meaning of "ought" has the kind of structure that it has on the Kratzerian view. Rather, what we want is the most simple and uniform metasemantic theory that can explain this and the other things that a metasemantic view should explain. For example, a plausible metasemantic theory should account for the data concerning normative disagreement. When someone accepts the sentence "We ought to ban Nazi symbols" and someone else accepts the sentence "We ought not to ban Nazi symbols," this usually constitutes a disagreement. Different metasemantic views face different challenges in explaining why this is so, and the difficulty of the relevant challenges does not track the degree of simplicity and unity that such views enjoy in relation to explaining the semantic structure of "ought." For instance, contextualist views, according to which ought-judgments express robustly representational beliefs about how things relate to certain contextually specified standards, may offer a very neat and simple explanation for why the meaning of "ought" would have the Kratzerian structure across different kind of uses, but struggle to accommodate the data concerning when we agree or disagree about normative issues (see, e.g., Finlay 2017).

In addition to the data concerning disagreement, the right metasemantic view also needs to explain whatever it is that needs explaining in relation to, for instance, the intuitions that fuel the open question argument, the relationship between normative judgment and motivation, and our plausibly conceptual knowledge of the supervenience of the normative on the descriptive (Venesmaa 2021). Undoubtedly there's much more that needs explaining. But this sample suffices to make it clear that we shouldn't be too quick to rule out the idea that capturing what needs to be captured by a metasemantic theory may require some complexity in the

metasemantics. This is not to say that we should make sacrifices with regard to simplicity and uniformity. We just need the most simple and unified theory that captures all the relevant phenomena.

This brings us naturally to Wodak's second objection to trying to occupy the space—one that we have proposed is available for an expressivist—of combining distinct metasemantic stories for practical and non-practical uses of normative terms with a unified semantic account.

> Second, consider how [the] non-expressivist explanation interacts with its expressivist counterpart. Here [the metasemantic expressivist] is committed to an unexplained coincidence. There is one *explanandum*: that the word "ought" means *Z*. There are two radically different *explanantia*; the expressivist, after all, is emphatic about the differences between representational beliefs and non-representational conative states. If the *explanantia* are radically different, why is the *explanandum* exactly the same? Why don't the radical differences between the states that we are expressing translate to differences in meaning? And, relatedly, why would we employ one word to express such radically different mental states? (Wodak 2017, 284–85)

This second objection from Wodak doesn't concern the complexity of the expressivist's metasemantic account as such. Instead, the worry is that the expressivist's account leaves a striking coincidence completely unexplained. In order to appreciate the force of this objection, it is helpful to first consider an example of a very simple expressivist view.

According to what we may call *simple expressivism*, "ought"-sentences express desire-like states of mind that are quite different from the belief-like states of mind expressed by descriptive sentences. So, whereas a sentence such as "The use of Swastika by the Finnish Air Force is not historically unrelated to the use of Nazi symbols" expresses a belief, the job of which is to describe and match the way the world is, a sentence such as "We ought to ban Nazi symbols" expresses a different kind of state of mind, perhaps opposition to allowing the use of Nazi symbols.

We may now try combining this simple expressivist view with the attempt to explain the same semantic structure on the basis of different metasemantic accounts of practical and

non-practical uses of normative language. If we do this, we end up being committed to the idea that even though the practical and the non-practical uses of "ought" express radically different states of mind, both kind of uses involve the very same meaning for "ought," where this uniform meaning is supposedly explained on the basis of this meaning somehow mirroring the structure of the very different mental states expressed by these sentences. Again, the problem, here, is not that the dualist metasemantics would be too complex. Rather, the problem is that the idea of a dualist metasemantics that appeals to very different types of states of mind and yet yields a completely unified semantic structure for "ought" commits one to an unexplained coincidence. An acceptable metasemantic theory doesn't tolerate this.

It's a good problem. It seems like a devastating problem for simple expressivism. There is no hope for an expressivist metasemantic view, such that would avoid the commitment to an unacceptable, unexplained coincidence, unless more structure—more structure suitable for being mirrored by the semantics of "ought"—is introduced in the expressivist's account of the states expressed by normative sentences. This is a very interesting result. But there is a further interesting result that can be obtained here, namely, that there is a brand of expressivism that plausibly has the resources for providing the kind of structure that is needed. We next present an expressivist view that has this nice feature: relational expressivism.

4. Relational expressivism and the problem of unexplained metasemantic coincidence

Let us suppose that Alex and Blue both accept that Nazi symbols ought to be banned. Here's what we believe is a plausible idea: Alex's and Blue's accepting that Nazi symbols ought to be banned is, very roughly, a matter of their being opposed to actions that have this or that property—who knows which one or which ones—and of their believing that failing to ban Nazi symbols has a relevant property (whatever property it is). Or perhaps we could say that Alex's and Blue's holding their view regarding the banning of Nazi symbols is, very roughly, a matter of their being opposed to

failing to ban Nazi symbols on the grounds of such failure's having some property that they would treat as relevant.

Alex and Blue may have very different normative perspectives. Perhaps Alex is a utilitarian who thinks that failing to ban Nazi symbols doesn't maximize happiness, and who is therefore opposed to failing to ban them. Perhaps Blue is a Kantian who thinks that failing to ban Nazi symbols is not compatible with the Categorical Imperative, and who is therefore opposed to not banning them. Or perhaps Alex and Blue are normal human beings and neither is very articulate about what their respective normative perspectives are like. Maybe Alex is opposed to things that are vaguely such and such—like *that* (mentally pointing, so to speak, toward *these* actions and policies), whereas Blue is opposed to things that are vaguely thus and so—like *that* (mentally pointing, so to speak, toward *those* actions and policies instead of these) (for appeal to the mental demonstratives of this sort, see Ridge 2014).

In any case, Alex and Blue are both opposed to things that have some—these or those—properties, and they both believe that failing to ban Nazi symbols has a relevant property. Even though their normative perspectives differ, they share an interesting similarity. They both are opposed to some type of actions—these or those—and believe, of the property that grounds their attitudes of opposition, respectively, that failing to ban Nazi symbols has that property. Their desire-like states of opposition and their suitably related representational beliefs (concerning Nazi symbols, in this instance) are related in the same way. They both are in the very same type of *relational state*, we may say, where this relational state is multiply realizable and differently realized by having some such desire-like state and a representational belief that are related in the relevant way.

As noted, we think that it is plausible to think that Alex's and Blue's holding the normative views that they hold is, very roughly, a matter of their being like this—a matter of their sharing this type of relational state. We also find it plausible that the sentence "We ought to ban Nazi symbols" expresses a relational state of roughly this kind. According to relational expressivism, normative sentences express states of this kind (see Schroeder 2013, Toppinen 2013, Ridge 2014).

Relational expressivism holds promise with regard to providing a suitably structured account of the states expressed by normative sentences—one that has the right kind of structure for the Kratzerian semantics to mirror. Above, we have characterized this kind of semantics as follows:

> The meaning of "ought" is a function from a proposition that gives the semantic value *true* just in case the proposition is true in all of the worlds that are ranked highest by standards of a certain sort, among a set of worlds restricted in certain ways, where the relevant standards and restrictions are contextually determined.

So, for example, the sentence "We ought to ban Nazi symbols" would be true just in case, among a set of worlds restricted in certain ways (e.g., to those in which we are able to ban Nazi symbols), in all of the worlds that are ranked highest by some relevant standards, we ban Nazi symbols. How would this kind of semantic structure mirror the structure of the mental states that are, according to relational expressivism, expressed by "ought"-sentences?

There are two options that a relational expressivist might pursue here. First, it is quite plausible that the representational beliefs that partly realize the relational states of mind expressed by normative sentences (e.g., "ought"-sentences) are beliefs concerning standards. When we consider the nonpractical uses of "ought"-sentences (about the requirements of etiquette, say), it is very natural to think that such sentences express representational beliefs about what is required (etc.) by certain standards. The expressivist may simply propose that this is a part of what's going on in the case of the practical uses, too.

Some expressivists have adopted this kind of idea. Ridge (2014, Ch. 1), for instance, suggests that claims about what's good express states that involve beliefs about what is highly ranked by certain standards. Likewise, claims about what ought to be done or about what must be done express states that involve beliefs about what is recommended or required by certain standards, and so on. That something like this is correct is, again, quite plausible. When we judge something to be good, it seems that we may always ask "By what standards?" When we classify actions as the ones that ought or

must be performed, we are committed to there being some grounds for why these are the actions that ought to, or must, be performed. And it is natural to think that we are thereby committed to there being some standards that help to articulate the relevant grounds.[8]

Above, we have suggested that when Alex and Blue accept that we ought to ban the use of Nazi symbols, this is a matter of their being opposed to things that have some—these or those—properties, and of their believing that failing to ban Nazi symbols has a relevant property. Perhaps, we said, Alex is a utilitarian who thinks that failing to ban Nazi symbols doesn't maximize happiness, and who is therefore opposed to failing to ban their use. And perhaps, we said, Blue is a Kantian who thinks that failing to ban Nazi symbols is not compatible with the Categorical Imperative, and who is therefore opposed to not banning them. Assuming that this is so, we may now understand this, somewhat more specifically, as follows: when Alex accepts that we ought to ban the use of Nazi symbols, this is a matter of their being opposed to actions that are not in accordance with the utilitarian standard (which requires that we maximize happiness), and of their believing that failing to ban Nazi symbols is not in accordance with this standard; when Blue accepts that we ought to ban the use of Nazi symbols, this is a matter of their being

[8] Does this rule out some forms of *particularism* that question the centrality of standards or principles to normative thought and talk? We think that particularists should agree that we are, in normative thought and talk, committed to the distribution of normative properties necessarily being determined by some necessarily true descriptive-to-normative principles, even if such principles are of no epistemic help. However, if the expressivist does need the standards to play a more ambitious role in normative thinking, then this may constitute a conflict with some interesting forms of particularism (for discussion on particularism, see, e.g., Dancy 2004, McKeever & Ridge 2006). Ridge (2014, 43–44) notes this issue in the context of his own favored formulation of a relational expressivist view and makes the interesting point that given that the appeal to standards is motivated by "quite general considerations in semantics for words like 'ought' that cut across both normative and non-normative uses," this "puts pressure on the particularist to offer alternative unified semantics or defend a kind of ambiguous view of words like 'ought'," where neither of these moves seems "terribly promising."

opposed to actions that are not compatible with the Kantian standard (which requires that we act in accordance with the Categorical Imperative), and of their believing that failing to ban the use of Nazi symbols is not in accordance with that standard.

Or consider the option of understanding Alex and Blue, more realistically, as being somewhat inarticulate about which properties are normatively relevant, or about what standards they endorse and reject. This, too, may be understood in terms of their endorsing or opposing standards of a certain kind. When Alex accepts that we ought to ban Nazi symbols, this can be understood to be, in part, a matter of their being opposed to actions that are *like that*, where having the relevant property (that is, being "like that") amounts to being incompatible with certain standards. Which ones? Well, those that rule out, for instance, actions "like that." In Alex's judgment that we ought to ban Nazi symbols this state of opposition then combines with a belief that failing to ban Nazi symbols has the relevant property—that is, with their belief that failing to ban Nazi symbols it not compatible with standards such that rule out, for example, actions "like that" (whatever actions they think of as being "like that").

In any case, no matter to what extent Alex and Blue are articulate about the standards that they have adopted, their accepting that we ought to ban Nazi symbols will be a matter of their being opposed to actions that do not meet standards of a certain kind, and of their believing that failing to ban Nazi symbols is not in accordance with the relevant sort of standards.

Above, we noted that there are two ways in which one might suggest, in line with relational expressivism, that the semantics of "ought" mirrors the states expressed by "ought"-sentences. The first was to appeal to the idea that "ought"-sentences express relational states that are always realized, in part, by representational beliefs concerning what is recommended or required by some standards. The second option that could perhaps be pursued here is the following. Instead of saying that the relational states expressed by normative sentences are always realized by beliefs that provide the structure for the Kratzerian semantics to mirror, one could suggest that the relational states themselves provide

the relevant structure. The idea would be that while the representational beliefs that partly realize these relational states need not concern standards, the relational states that they partly realize do so. Let us return, for example, to our first characterizations of the relational expressivist idea. We first proposed that when Blue accepts that we ought to ban the use of Nazi symbols, this could be construed, roughly, as a matter of their being in a state of being opposed to actions that have some particular property—that of being incompatible with the Categorical Imperative, say—and of believing that failing to ban the use of Nazi symbols has that property. Perhaps one could now propose that being opposed to actions that have some particular property amounts to acceptance of a standard. And perhaps one could then propose that if one now, in addition to accepting some standard, has a belief to the effect that some action has some property that suitably relates to the standard in question (e.g., has a property such that actions with that property are ruled out by the standard), then one's states of opposition and belief will be related to each other in a way that constitutes a judgment the content of which concerns a standard. In this way, even if the representational belief that partly realizes the relational state expressed by a normative sentence would not concern the relation of anything to any standard, the relational state itself could be taken to have the functional profile of a judgment or a belief that does concern such things. We shall not say more about this kind of idea here. But this kind of option may nevertheless be worth keeping in mind.

If the kind of relational expressivist metasemantics outlined above is correct, then it seems unsurprising that the practical uses of "ought"-sentences have the kind of semantics that we have assumed they have. The semantics now nicely mirrors the structure of the states of mind expressed by "ought"-sentences. We haven't said anything about *how*, exactly, the account of the states expressed by the relevant sentences explains the semantics. The talk of "mirroring" isn't perhaps very satisfying, ultimately. However, this seems acceptable in the present context. We have been operating here with the assumption that the broader expressivist project of explaining meaning in terms of the states of mind that sentences express is workable. This idea can of course be contest-

ed. However, in the present context, we have not been concerned with defending this idea. And the problems we have been addressing—the problem of diverse uses, the problem of disunified metasemantics, and the problem of metasemantic coincidence—are not supposed to be problems for this broader project of explaining the meaning of sentences in terms of the states of mind that they express. Instead, these latter problems have been raised for expressivism against the background of assuming, at least for the sake of the argument, the legitimacy of the expressivist metasemantic project.

In the context of this project, we may assume that if "Snow is white" expresses a belief that snow is white, this explains the meaning, or the truth-conditions, of the sentence. Likewise, in the context of this project, we may assume that if "One ought not to eat peas with a spoon" expresses a belief about what is required by certain standards, this explains the meaning, or the truth-conditions, of the sentence. That is, we may assume that this explains why the sentence (as uttered in an appropriate context) is true just in case peas are left uneaten in all the worlds restricted by certain background conditions and ranked highest by the standards of etiquette. If this is all correct, and if "We ought to ban the use of Nazi symbols" turns out to express the kind of relational state that this sentence expresses, according to relational expressivism, then this plausibly explains why this sentence, too, has a meaning similar to that of "One ought not to eat peas with a spoon." That is, the relational expressivist account explains why this sentence, too, says of something—of banning the use of Nazi symbols—that that is what is done in all the worlds the set of which is restricted in certain ways and ranked in accordance with some suitable standards.

It's worth emphasizing that, according to relational expressivism, when the sentence "We ought to ban the use of Nazi symbols" is used in a practical role, it does not *express* any representational belief concerning standards. Rather, the sentence expresses a relational state that is differently realized by different desire-like states and representational beliefs in different contexts of use. So, when Alex the utilitarian accepts "We ought to ban the use of Nazi symbols," the sentence, as used in the relevant context, does not express a be-

lief that failing to ban the use of Nazi symbols is not in accordance with the utilitarian standards. These standards need not be, and are likely not to be, contextually specified, in this instance, in the way that the standards of etiquette might be specified in a context in which "One ought not to eat peas with a spoon" is used. What is contextually specified is just that what's in play are the standards relevant to the practical use of "ought," the standards of practical reason, we might say (with Ridge 2014), or of what to do—the "genuinely" or "robustly" normative standards. Whether these standards are utilitarian, or Kantian, or something completely different, is left for practical deliberation or normative theorizing to decide.

5. Conclusion

Expressivists wish to explain the meaning of (some of) the normative language by appealing to the practical role that such language plays. We have here addressed a problem for the expressivist proposal, *the problem of diverse uses*, which arises from the fact that normative terms often figure in sentences that do not play any interestingly practical role. The challenge is that of explaining, in the face of this fact, the meaning of normative language in a sufficiently unified manner.

We have proposed, as the first step toward responding to this challenge, that expressivism should be understood as a view in metasemantics. Expressivists may then suggest that they can make sense of normative terms as having a unified meaning across the practical and non-practical uses of normative language. It's just that this unified meaning is sometimes explained by the practical role of normative language, and sometimes by its non-practical, representational role.

This kind of move gives rise to two further problems. First, *the problem of disunified metasemantics* draws our attention to the fact that the expressivist metasemantics is somewhat more complex and disunified than some of its alternatives. We have granted that we should seek a metasemantics for normative language that is as simple and unified as is possible, given that it allows us to explain all the semantic data that requires explanation. However, the practical uses of

normative language do differ, in many ways, from the non-practical uses. And some of the differences plausibly have to do with the meaning of the language used. We have suggested that the phenomena that are relevant here—the disagreement data, "open question" intuitions, etc.—may very well justify, or indeed require, some complexity in the metasemantics. We believe that expressivism is well-placed to capture complexity of the relevant kind, but determining whether or not this is so is beyond the scope of this discussion.

Second, though, *the problem of unexplained metasemantic coincidence* also needs to be addressed. Given the expressivist idea that the practical and non-practical uses of normative language express importantly different kind of states of mind, the expressivist account would seem to be committed to it being completely coincidental, and wholly unexplainable, that normative terms should have the very same core meaning across both practical and non-practical uses. We have granted that this seems like a devastating problem for a simple expressivist view, according to which normative sentences sometimes express representational beliefs (about, say, the requirements of the norms of etiquette), and sometimes desire-like attitudes of a wholly different sort (disapproval of the use of Nazi symbols, for example). However, there is no unexplained coincidence in how the meaning of normative terms is explained across their different uses, given the truth of one kind of expressivist view, relational expressivism. According to this view, normative sentences always express states of mind that involve representational beliefs relating things to certain standards. This offers promise with regard to providing us with a metasemantics that has a unified enough structure for a unified semantics to mirror.

Is there still some question that we would have alluded to, but that would have been left unaddressed? One of Wodak's worries concerns the expressivist's resources with regard to explaining why we would employ *one word* to express radically different mental states. Given the relational expressivist view, we may now replace "radically different" with "somewhat different." Still, one might wonder why it should be that judgments involving the term "ought," or a word that "ought" translates to, would express two different kinds of judgments, practical and non-practical, about what is re-

quired or recommended by certain standards. We have not directly addressed this particular question. But as we see it, given the metasemantics for "ought" roughly outlined in the previous section, there shouldn't be anything terribly surprising or mysterious about the fact that we use one word for expressing judgments or mental states of a somewhat different kind. The explanation for why we would have a word for expressing the relevant kind of mental states plausibly has to do with the usefulness, or necessity, of standards in the guidance of action and attitude management. Plausibly, the practical uses of "ought" have *priority* over the non-practical uses, in that the idea of a community that would only use "ought" in non-committal or non-practical ways seems very strange, while the converse doesn't seem to hold. We need the practical or committal oughts to guide and coordinate our actions and attitudes. But we are also bound to have an interest in the ways in which those surrounding us guide and coordinate their actions and attitudes, even if their commitments differ from those of ours. And we are also bound to have an interest in the various possible ways of guiding and coordinating actions and attitudes. It would make perfect sense, then, to use the same words for relating things to various standards for choice and belief (etc.), regardless of whether the relevant language would have a directly practical use for us, or instead be used in tracking some "standard-involving" facts in the absence of a direct practical concern.

Even supposing that we have provided satisfying responses to the problems of diverse use, disunified metasemantics, and unexplained metasemantic coincidence, much more work remains to be done on related issues. For example, it would be nice to have an account of how, exactly, the expressivist metasemantics explains semantics. Also, it remains to be determined (as far as we know) what the correct semantics for the various normative terms is like. We have only toyed with one toy view regarding the meaning of one normative term ("ought"). Plausibly, we will only be in a position to figure out what the expressivist explanation for the meaning of various normative terms looks like once we know what the right semantics is for these terms. However, the relational expressivist view allows for a lot of variation in how, exactly, the relational states expressed by normative sentenc-

es should be understood. This makes us optimistic that whatever the right semantics for normative language is going to be, relational expressivism will provide interesting resources for finding the right kind of explanation for it. At the very least the challenges that we have here addressed give us no reason to be skeptical about the prospects of this brand of expressivism.[9]

Tampere University
University of Helsinki

References

Ayer, A. J. (1936). *Language, Truth, and Logic* (2nd edition, 1946). London: Gollancz.

Bex-Priestley, G. (forthcoming). "Expressivists Should Be Reductive Naturalists." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*. Oxford: Oxford University Press.

Blackburn, S. (1998). *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press.

Bronfman, A. & J. Dowel (2018). "The Language of 'Ought,' and Reasons." In D. Star (ed.), *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.

Carr, J. (2018). "Deontic Modals." In T. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics*. New York: Routledge.

Chrisman, M. (2016). *The Meaning of 'Ought': Beyond Descriptivism and Expressivism in Metaethics*. New York: Oxford University Press.

Dancy, J. (2004). *Ethics without Principles*. Oxford: Oxford University Press.

Finlay, S. (2014). *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

Finlay, S. (2017). "Disagreement Lost and Found." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 12*. Oxford: Oxford University Press.

Fletcher, G. (2014). "Moral Utterances, Attitude Expression, and Implicature." In G. Fletcher & M. Ridge (eds.), *Having It Both Ways: Hybrid Theories and Modern Metaethics*. Oxford: Oxford University Press.

[9] We thank an anonymous reviewer for a helpful set of comments on an earlier version of the paper.

Gibbard, A. (2003). *Thinking How to Live*. Cambridge, Mass.: Harvard University Press.

Kratzer, A. (2012). *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin.

McKeever, S. & M. Ridge (2006). *Principled Ethics: Generalism as a Regulative Ideal*. Oxford: Oxford University Press.

Ridge, M. (2014). *Impassioned Belief*. Oxford: Oxford University Press.

Schroeder, M. (2008). *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.

Schroeder, M. (2010). *Noncognitivism in Ethics*. Abingdon: Routledge.

Schroeder, M. (2013). "Tempered Expressivism." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 8*. Oxford: Oxford University Press.

Sinclair, N. (2007). "Propositional Clothing and Belief." *The Philosophical Quarterly* 57, 342–362.

Sinclair, N. (2021). *Practical Expressivism: A Metaethical Theory*. Oxford: Oxford University Press.

Strandberg, C. (2011). "A Dual Aspect Account of Moral Language." *Philosophy and Phenomenological Research* 84, 87–122.

Thomson, J. J. (2008). *Normativity*. Chicago: Open Court.

Toppinen, T. (2013). "Believing in Expressivism." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 8*. Oxford: Oxford University Press.

Venesmaa, V. (2021). "Explaining Our Knowledge of Normative Supervenience." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 16*. Oxford: Oxford University Press.

Väyrynen, P. (2022). "Practical Commitment in Normative Discourse." *Journal of Ethics and Social Philosophy* 21, 175–208.

Wodak, D. (2017). "Expressivism and Varieties of Normativity." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 12*. Oxford: Oxford University Press.