METASEMANTICS, MODERATE INFLATIONISM, AND CORRESPONDENCE TRUTH

by

GRAHAM SETH MOORE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Philosophy)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Metasemantics, Moderate Inflationism, and Correspondence Truth

submitted by Graham Seth Moore in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

in Philosophy

Examining Committee:

Ori Simchen, Professor, Department of Philosophy, UBC

Supervisor

Roberta Ballarin, Associate Professor, Department of Philosophy, UBC

Supervisory Committee Member

Matthew Bedke, Professor, Department of Philosophy, UBC

Supervisory Committee Member

Carrie Jenkins, Professor, Department of Philosophy, UBC

University Examiner

Hotze Rullmann, Associate Professor, Department of Linguistics, UBC

University Examiner

Gila Sher, Professor, Department of Philosophy, University of California, San Diego

External Examiner

Abstract

An object-based correspondence theory of truth holds that a truth-bearer is true whenever its truth conditions are met by objects and their properties. In order to develop such a view, the principal task is to explain how truth-bearers become endowed with their truth conditions. Modern versions of the correspondence theory see this project as the synthesis of two theoretical endeavours: *basic metasemantics* and *compositional semantics. Basic metasemantics* is the theory of how simple, meaningful items (e.g. names and concepts) are endowed with their contributions to truth conditions, and *compositional semantics* is the theory of how the meanings of simple items compose to generate (among other things) the truth conditions of sentences.

Understanding truth along these broad lines was once popular; it was first championed by Field (1972). However, the once-popular conception of its tasks included an over-ambitious view of basic metasemantics. It was thought that reference needed to be *analyzed* (or *reduced* a posteriori) in terms of more fundamental, non-semantic relations (e.g. causal relations, indication, or teleological relations, in the case of mental representation). Obstacles in providing such an analysis engendered skepticism towards this understanding of truth and eventually gave way to its deflationary competitors.

This dissertation aims to defend the modern, object-based correspondence theory against its rivals—especially deflationism. Chapter one provides a historically-grounded overview of the theory. Chapter two identifies two points of contrast between the correspondence theorist and the deflationist: they employ different orders of explanation for the variety of semantic phenomena, and they (traditionally) take different attitudes towards the prospects of reduction. Situating the dialectic in this way allows me to develop a middle ground: a moderate version of inflationism that takes the inflationary explanatory structure and combines it with a non-reductive, pluralist approach to basic metasemantics. Chapter three expands on the details of this pluralist account of reference. Chapter four contrasts the view with another rival approach to basic metasemantics: metasemantic interpretationism. And finally, chapter five applies the theory to answer another question of broad philosophical interest: *what role does our conception of truth play in inquiry about the world?*

Lay Summary

An object-based correspondence theory of truth holds that truth is entirely explicable as the product of (i) how words represent things, (ii) how the meanings of sentences are determined by the meanings of words, and (iii) how things stand in the world. This view was once popular, but despite its many virtues, it fell out of favour because its advocates overreached in their theoretical ambitions with regards to (i). In its stead, a rival view emerged which sought to trivialize the theory of truth.

This dissertation aims to defend the object-based correspondence theory against its rivals. The key is to explain how (i) can be theorized in a way that is non-trivial yet realistic. Following this, the theory is applied to address the question of how thinking in terms of truth aids our investigation of the world.

Preface

This thesis consists in original, independent work by Graham S. Moore. Chapters one, three, four, and five are unpublished.

A condensed version of chapter two is published as ‘Between Deflationism and Inflationism: A Moderate View on Truth and Reference’ in *The Philosophical Quarterly* 72/3: 673–694. Some of this material also appears in chapter one. It is reprinted here with permission from Oxford University Press.

Table of Contents

Acknowledgements

There are many people to whom I owe a great debt of gratitude for their help in making this dissertation possible. First of all, I would like to thank my dissertation committee—Ori Simchen, Roberta Ballarin, and Matthew Bedke—for greatly expanding my philosophical horizons and for their helpful comments on drafts of this material. I owe additional thanks to Roberta Ballarin for also acting as a teaching mentor during my time at UBC. Finally, I would also like to thank Jonathan Jenkins Ichikawa for mentoring me during my first few years in the program.

Of course, none of this would have been possible without my wife (and fellow doctoral student) Kelsey Vicars. Thank you for reading and commenting on all of this material. Thank you also to Daniel Saunders for feedback on chapter two.

Finally, I would like to thank SSHRC for their financial support during my first three years of the program.

Chapter 1: The correspondence theory as a project in semantics and metasemantics

1.1 Introduction

The so-called ‘correspondence theory of truth’ has a long history of being expressed in platitudes. In the beginning (or rather, in the beginning of philosophy as we know it), Aristotle famously wrote that “To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true” (*Metaphysics:* 1011b25). Sometime later, St. Thomas Aquinas claimed that truth consisted in conformity between things and the intellect (*De veritate*: q. 1, a. 1, 5.162–6.200). And then, later still, in what is almost recent memory, Bertrand Russell and G.E. Moore proclaimed that truth is a matter of beliefs corresponding with the facts (Russell 1997; Moore 1953). In their classical contexts, these platitudes were always expressed to signal rejection of some other more unpalatable view. These early correspondence theorists would propound such things in order to distance themselves from the idea that truth is in the eye of the beholder, or a mere matter of opinion, or whatever is most useful to believe, or whatever coheres with the norms of inquiry.

Each of these platitudes gives voice to the same basic thought shared by those who travel under the ‘correspondence theory’ label: that truth is a matter of accurately representing how things are in the world. Traditionally, the philosophers who attempted to develop this thought have looked to metaphysics and ontology. They would build upon the correspondence intuition by developing metaphysical theories of the nature of propositions and facts. But as time went on, the efforts to explicate the correspondence intuition evolved in a direction that drifts away from traditional metaphysics. This is because a series of developments in logic, linguistics, semantics, cognitive science, and philosophy now jointly inform our modern understanding of truth. At least, that is the story that I would like to tell. It is thus my aim for this chapter to tell some of this story and trace some of this history. Ultimately, I would like to outline a modern version of the correspondence theory that is sensitive to these developments.

Since our concern is a theory of truth, we must begin with a word on the questions these theories attempt to answer. The concern of a correspondence theory is with the metaphysics of truth. The aim is to say what the nature of truth is and provide the sorts of explanations that can be given for when something is true. Suppose I pick up a shiny piece of rock from a river bed and display it in my hand for everyone in my vicinity to see. I then pronounce ‘*this* is a piece of gold’. According to our intuitive, folk understanding of language, my pronouncement may be true or false, depending on various factors. Intuitively*,* it depends on what the sentence *means* and how things stand *in the world*—whether that particular rock in my hand is, in fact, a piece of gold. This reflects the general insight of the correspondence theory: that the truth of what I say depends on the worldly states that I represent. This intuition is fine as far as it goes, but the hard problem for the correspondence theorist is to develop it into something revelatory and systematic.

Broadly speaking, everyone who writes on this topic can agree on this much: truth is a property borne by some sort of truth-bearing item—a proposition or a thought or a sentence—generically called a *truth-bearer.* The truth-bearers are each associated with conditions for their truth (and falsity); these are called *truth conditions* (and falsity conditions). They are *true* when their truth conditions are met and are false when they are not met (or when their falsity conditions are met).

Although this may not sound like much, it at least points us in the direction of the further questions to be asked. For one, we need to ask *what the truth-bearers are.* When I say '*this* is gold', is the true thing (or false thing) *my utterance* (a particular act of communication)? Or is it *the thought* I express with my utterance (a particular mental entity)? Or is it something more abstract, like a *proposition* (the thing my sentence expresses as its *meaning*)? For the second cluster of questions, we also need to ask what the truth conditions are—what makes a truth-bearer *true*? The generic answer from the correspondence theorist is that it has something to do with how things stand in the world. A truth-bearer is true if it *corresponds to reality,* as the saying goes. But as we will see, this needs to be precisified and systematized. Since the truth-bearers can come in all sorts of varieties and orders of complexity, we need a theory of truth conditions that reflects this variety. Finally, even if we can answer the previous two questions, there may still be a further explanatory task. We may also need to explain *why* the truth-bearers are endowed with their particular truth conditions. Why, for instance, does my utterance of '*this* is gold' mean what it does, so that it is true if and only if the rock I hold is an instance of gold? Addressing these three clusters of questions is what I take to be required to 'explain the nature of truth'.

1.2 The fact-based correspondence theory

Let's begin the history. For our purposes, it is appropriate to start with the advent of analytic philosophy. (We could have begun with earlier philosophers or other traditions, but then there are endless possibilities. It makes sense to start with the tradition that engendered the developments we'll discuss shortly.) In that case, our first protagonists are two of the founders of analytic philosophy, G.E. Moore and Bertrand Russell. At one stage during their philosophical development, as part of their break with the earlier idealist traditions, Moore and Russell each advocated for what is now known as the classical *fact*-based correspondence theory of truth.[1] Here are their own statements of their views:

> To say that this belief is true is to say that there is in the universe *a* fact to which it corresponds; and to say that it is false is to say that there is *not* in the universe any fact to which it corresponds. (Moore 1953: 277)

> A belief is true when there is a corresponding fact, and it is false when there is no corresponding fact. (Russell 1997: 129)

Each of these can be thought of as a natural way to give voice to the intuition that truth is a matter of representing reality. When I believe that the rock I hold in my hand is a piece of gold, my belief is true if it accurately represents the relevant portion of reality concerning the rock in my hand—that is, if it *corresponds* to a fact.

The first thing to notice about these definitions is that, for both Moore and Russell, a truth always requires the *existence* of a corresponding fact. In every instance where something is true, it is because there *exists* something (the fact) that makes it true. This is one of the defining features of the classical correspondence theory: truth is explained by postulating a local ontology of facts; the facts must exist to witness the truths. Moore and Russell thus explain truth by giving an *ontological* theory, and the ontology is one that posits *facts.*

Since the theory gets its mileage out of its ontology of facts, it invites further questions about how to understand these facts. To these questions, both Moore and Russell give roughly the same answers.[2] For them, facts are understood as a kind of complex structured entity that is composed of both particular objects and abstract universals (see e.g. Moore 1899; Russell 1997: 129). For example, the fact that I represent when I say '*this* is a piece of gold' will be a complex entity composed of the rock in my hand and the property of *being a piece of gold.* For lack of a better way to talk about these things, we might represent this fact as the ordered pair <*this rock, being a piece of gold*>. Moreover, the *existence* of this thing is supposed to explain why it is true that *this is a piece of gold.* This is why it is apt to call this the *fact-based* correspondence theory.

The second component of the fact-based correspondence theory concerns the *truth-bearers*—the things that bear truth and falsity. In the quotations above, both Moore and Russell take *beliefs* as the bearers of truth. Each of them had their own peculiar reasons for this (see e.g.

[1] This was not their first stopping-point in their thinking about truth. In his (1899), Moore advocates for an identity theory of truth, whereby a true proposition is taken to be identical to a fact.

[2] They gave the same answers *at one time or another*. However, both Moore and Russell's views evolved throughout their advocacy of correspondence-based truth; so it is misleading to present any timeslice of either Moore or Russell's views as if it were official.

Russell 1997: ch. XII).[3] But instead of attending to the idiosyncrasies of this particular temporal stage of Moore and Russell's views, it is better to present the fact-based correspondence theory in its more general form. In that case, the most typical candidates for the theory's truth-bearers are *sentences* (of either public or mentalistic languages) and *the objects of belief*: i.e. *propositions.*

Should a theory choose propositions as the primary truth-bearer, it would owe an account of what propositions are. The standard answer is that propositions are *the objects of belief* (and other propositional attitudes), the *meanings of declarative sentences* (taken in context), and the bearers of correspondence relations and hence truth and falsity (King 2014a).[4] But besides this, there is no further non-controversial answer as to what else propositions are (see King, Soames & Speaks 2014 for a recent contribution to this debate). Any acceptable 'propositionalist' version of the correspondence theory would be required to stake a position in this perennial controversy.

The third and final component of the fact-based correspondence theory has to do with the *correspondence* relation that is borne between the true propositions and the facts. Proponents of this view need to explain what this is. What, after all, does it mean for a proposition to *correspond* with a fact? When I say that *this is a piece of gold,* what relation does my statement bear to the gold in my hand and the facts that surround it?

The most common line of response appeals to the details about the alleged nature of propositions. Typically, the correspondence theorist will posit that propositions have an internal structure (see David 2018; Rasmussen 2014). Perhaps the proposition that *this is a piece of gold* has a structure that reflects the subject-predicate form of the sentence, and it has a constituent that represents the rock in my hand and a constituent that represents the property of being gold.[5] We may thus represent the proposition as the ordered pair, <*representation of the rock, representation of the property of being gold*>. Once we have come this far, we can then understand the correspondence between a true proposition and a fact as a kind of isomorphism between proposition and fact. The idea is that a true proposition will match its corresponding fact constituent-for-constituent and structure-for-structure. So, for instance, the proposition that *this is a piece of gold* is *true* because there exists a fact that is mirrored by the constituents and the structure of this proposition. Thus, the full explanation of truth, according to this version of the

[3] Russell was concerned about the nature of false propositions and the 'unity of the proposition', so by the time that he advocated the correspondence theory, he took *beliefs* to be the truth-bearers, and analyzed belief according to his multiple relations theory.

[4] This should be taken as a theoretical definition; propositions are the entities that play these three roles (if there are any such entities). Even though this suggests that propositions are theoretical entities, they are still taken (by the propositionalist picture) to be the objects of common (i.e. non-theoretical beliefs) and the meanings of everyday (i.e. non-theoretical) sentences.

[5] There is a substantive point of controversy as to what these constituents *are*—whether they should be construed along Fregean or Russellian lines. The classic contributions to this dispute are Frege (1956) and Russell (2020). See David (2018) for a discussion of how this issue bears on the present point.

correspondence theory, will appeal to the existence of a structured fact, which is mirrored by a structured proposition.

1.3 Criticisms of the fact-based theory & logical atomism

If any of this sounds appealing so far, that may be because we have only attended to a relatively simple truth of subject-predicate form. When we consider truths with more complicated structures, the view becomes rather unwieldy.

Suppose that the rock in my hand is gold, not copper. Now consider the truth that *either this rock in my hand is gold or it is copper.* What fact would make this proposition true? According to the fact-based account, there must be some complex fact that makes this proposition true: namely, the fact that *either this rock in my hand is gold or it is copper.* We can likewise consider truths of conjunctive form, conditional form, and subjunctive form. The fact-based correspondence theory would explain these by postulating the existence of conjunctive facts, conditional facts, and subjunctive facts. For any complicated truth you can imagine, this theory must posit the existence of an equally complex fact.

However, it has struck many philosophers that postulating these complex facts is, at best, unnecessary and, at worst, obfuscating. To explain the truth that *either this rock in my hand is gold or it is copper,* it would seem that (at most) I only need the fact that the rock in my hand is gold. Ludwig Wittgenstein expressed much the same sentiment:

> Whatever corresponds in reality to compound propositions must not be more than what corresponds to their several atomic propositions. (Wittgenstein 1961: 100)

What's more, in order for these complex facts to exist, there must be constituents corresponding to the logical terms 'or', 'and', and 'not'. Some philosophers have found this preposterous. When Russell became disillusioned with his earlier version of the correspondence theory, he wrote:

> You must not look about the real world for an object which you can call "or", and say "Now look at this. This is 'or'". (Russell 2007: 209–10)

Besides these two complaints, I would also add that the explanations offered by the fact-based correspondence theory end up being fairly shallow. Consider again how it explains the truth of an arbitrary proposition. Whatever syntactic form the proposition takes, the fact-based theory will spit back a fact that matches its structure and claim that this fact is the explanation of its truth.

Again, why is it true that *either this rock in my hand is gold or it is copper?* The fact-based theory answers that *it's a fact* that *either this rock in my hand is gold or it is copper.* But the theory never attempts to explain *how* the structure of this proposition affects its truth conditions. It merely reifies the structure of the proposition into the entity that's supposed to give the explanation—namely, the fact.

Each of these criticisms has been much discussed in the literature, and I do not claim that any of them is decisive (see David 2018). Nonetheless, they expose the weak points of the theory and point in the direction in which it can be improved.

After abandoning the pure fact-based theory, Bertrand Russell, along with the early Wittgenstein, advocated for a mixed view known as logical atomism (Russell 2007; Wittgenstein 2001). The basic tenet of this theory is that the only existent facts are the logically simple ones.[6] These explain the truth of all the atomic propositions (*viz.,* the propositions containing no connectives).[7] Then, for the complex propositions, their truth values are explained by the operation of the truth functions of the connectives on the truth values of the atomic propositions. Take again my example of the truth that *either this rock in my hand is a piece of gold or it is copper.* The truth of the first disjunct, *this rock in my hand is a piece of gold,* is explained by the existence of a corresponding fact (assuming that it is atomic; see fn. 6). The other disjunct, which says that *this rock is copper*, is false because it doesn't correspond to any fact.[8] Following this, the truth of the complex proposition is explained by the first disjunct being true and the truth function associated with the connective 'or'. So, according to logical atomism, truth is not to be understood as correspondence in all cases. The theory is pluralistic about the nature of truth. Truth is only correspondence to the facts for the atomic propositions. It is explained recursively for complex propositions by the truth values of the atomic propositions and the semantic operations of the logical connectives.

By renouncing the ontology of complex facts, logical atomism gathers several distinct advantages over the pure fact-based theory. For one, it excises some of the ontological excesses of the pure fact-based theory, so it scores the advantage of parsimony by point of comparison. In

[6] For each of these authors, the apparently simple sentence 'this rock in my hand is a piece of gold' would not count as simple upon analysis. The reasons for this stem from their respective epistemological and metaphysical commitments, which we need not concerns ourselves with here.

[7] Russell's version of the theory includes both positive and negative facts (e.g. *that Socrates is dead* and *that Socrates is not alive*) and universal and existential facts (e.g. *that all men are mortal* and *that some men are mortal*) among the logically simple ones. He excludes facts that are conjunctive, disjunctive, and conditional. Moreover, he advocates for a correspondence theory (for the atomic propositions) that allows for *two* types of correspondence. The proposition 'Socrates is dead' corresponds *truly* to the fact that *Socrates is dead* and the proposition 'Socrates is alive' corresponds *falsely* to the fact that *Socrates is dead.* See lecture III and V of (2007).

Wittgenstein, on the other hand, did not countenance negative, universal, or existential facts. His version of logical atomism only includes positive ones concerning 'simple' objects, properties, and relations.

[8] According to Russell's view from (2007), the proposition is false because it does not *correspond truly* to any fact, and, moreover, it *corresponds falsely* to the fact that *the rock in my hand is gold (not copper).*

addition (and as a result), it also gains the benefit of improved explanatory depth. Unlike the pure fact-based theory, logical atomism actually explains how the connectives of propositional logic interact with truth conditions.

Logical atomism thus contains the seeds of a very powerful idea. The idea is that we can account for truth by essentially taking two steps. First, we explain how truth works for a collection of base cases (in this case, the atomic propositions), and then, we explain how truth works for logically complex propositions based on their simpler parts. In a word, the idea is to explain truth *by recursion.* Moreover, by giving the theory this overall recursive structure, we gain a level of systematicity that we wouldn't have otherwise achieved. With only the small collection of principles in our recursive explanation, we get a general explanation of how truth conditions are generated according to propositional structure.

That being said, there is a case to be made that logical atomism is only a half measure. This is because it retains the defects of the pure fact-based theory at the level of the atomic propositions and facts. For instance, it retains the tactic of explaining truth for atomic propositions by positing local ontology; it explains the truth that *this is a piece of gold* by positing an *entity* (the fact that *this is a piece of gold*). But as W.V.O. Quine later pointed out, this still has the air of a vacuous explanation:

> What on the part of true sentences is meant to correspond to what on the part of reality? … [P]erhaps we settle for a correspondence of whole sentences with *facts:* a sentence is true if it reports a fact. But here again we have fabricated substance for an empty doctrine. The world is full of things, variously related, but what, in addition to all that, are facts? They are projected from true sentences for the sake of correspondence. (Quine 1987: 213)

Quine speaks of sentences as the truth-bearers rather than propositions, but the point is still the same. A theory would be better off without relying on a postulated ontology of facts. Not only that, but logical atomism *only* ever endeavours to explain the truth-conditional contributions of the propositional connectives. It does not offer any explanation of how *sub-propositional* structure impacts truth conditions. For this reason, logical atomism still leaves much about the nature of truth conditions unexplained.

1.4 The object-based correspondence theory

To remedy these problems, we need a theory that attends to the inner workings of sub-propositional structure. This is where the modern object-based correspondence theory makes its entrance.

To begin, let's start with the ontological presuppositions. With a nod to Quine, the ontology of the theory starts with the objects we'd pre-theoretically take to inhabit the world: tables, chairs, rocks, people, and so on. Later on, we may need to add numbers, sets, properties, possible worlds, or even facts—*if* our metaphysics calls for it. But this would be settled on metaphysical grounds, not by the basic commitments of a theory of truth.

Let us return to our example to give a sense of the improved theory. In this example, I have uttered the sentence 'this is gold' while demonstrating a rock in my hand. Now, what would it take to explain the truth of this sentence token? (Notice that I am talking about the *sentence token* rather than the proposition that it expresses. The shift of the truth-bearer will be addressed later.)

Here is the suggestion. First, we explain the referential relations that the sentence's lexical parts bear to the worldly objects they represent. To this end, we must tell some story as to why my use of the demonstrative '*this*' refers, in this context, to the rock in my hand. (We will say more about this story later.) We must also tell some story about why my use of the predicate 'is gold' applies to all and only gold things. (Perhaps it is in virtue of expressing the property of *being gold*, but at this stage, we want to be neutral towards nominalism and realism about universals.) In doing so, we see each (significant) part of the sentence as a symbol that bears a referential relation to things. We can call the facts about these referential relations the *primitive semantic facts.*

Once the primitive semantic facts have been accounted for, the next step is to attend to the structure of the sentence. In this case, the sentence in question has a basic subject-predicate syntactic structure. We can say that the significance of this syntactic structure is that it carries with it a semantic rule. This rule tells us how to *combine* or *compose* the semantic contributions of the expressions in the subject and predicate position into something true or false. In effect, this rule tells us that the sentence attributes the property expressed by 'is gold' to the object referred to by 'this'; or, to put the same point in an ontologically-neutral way, the rule tells us that the sentence is true if, and only if, the object referred by 'this' satisfied the predicate 'is gold'. Once we combine these pieces, we get the result that the sentence is *true* if and only if *this is gold.* Since it is a piece of gold (let's suppose), the sentence is true. All in all, we have come to an explanation of the truth of this sentence.

It is worth repeating, for emphasis, the two big ideas that figure into this explanation. The first idea is that certain simple expressions—names, demonstratives, predicates, relational terms —stand in referential relations—that is, relations between symbols and the things they symbolize. Singular terms *refer* to things and predicates *apply* to things, but each may be called a *referential relation.* These relations are the subject matter of the theory of primitive semantics, which is also sometimes called the theory of reference. The second big idea is that there are rules for combining the semantic properties of simpler expressions into the truth conditions for sentences, according to the syntactic form of the sentence. These rules are the subject of the theory of *semantic composition.* The theory of truth that we get is essentially a synthesis of these two theories.

So then, what *is* truth, according to this theory? In short: a sentence (or any other truth-bearing item; we will return to this) is true if its truth conditions obtain, and its truth conditions are explained as a product of the theory of reference (how expressions relate to objects) and the theory of semantic composition (how the semantic contributions of subsentential expressions determine truth conditions according to sentential form). We can call this the modern *object-based* correspondence theory (the terminology is from Glanzberg 2015). We have already given a toy example of how this account proceeds in the simple case; the remaining task is to outline how it will work in full generality. But before we get into that, a few remarks are in order.

First, why call it the 'object-based correspondence theory'?[9] The reason to call it '*object*-based' is because it eschews the fact-based ontology of the classical theories of Moore and Russell. Unlike their views, this theory does not require the existence of facts; it just needs the existence of *objects* for the singular terms to refer to and the predicates to apply to.[10] Moreover, we call it a '*correspondence* theory' because it respects the basic intuition that truth is a matter of representing things *as they are.* According to this theory, truth is a matter of referring to things and then attributing to them the properties they have. Since this theory retains the fundamental insight that truth is a matter of representing portions of the world as they are, it deserves to count as a correspondence theory.

Second remark. I have stressed that this correspondence theory is the synthesis of two theories: the theory of reference (or primitive semantics) and the theory of compositional semantics. It is also worth stressing that each of these is a fairly open-ended topic in its own right. In order to give an overview of the current state and prospects of the correspondence theory, we must say a few words about both of them. But since they are large areas of research, the discussions must be brief summaries.

The final item on the agenda will concern the topic of truth-bearers. As I flagged, the

[9] There is some debate over whether this approach ought to count as a *bona fide* correspondence theory (David 2018: 255). This is for two reasons. First, this conception does not identify *truth* with any unifying property of sentences/propositions (e.g it does not claim that *truth is correspondence with the facts*). Rather, it claims that truth is just a matter of the sentence's semantically-generated truth conditions being satisfied. For this reason, we might want to take a page from Tarski (1944) and label this the 'semantic conception of truth'. In this author's opinion, the choice over terminology, on this basis, is an insignificant issue. Secondly, the debate over whether this ought to count as a correspondence theory ended up evolving into a debate over whether or not the basic reference relations are 'real' or 'robust'. If reference is given a deflationary reading, then the resulting theory would not generally be counted as a correspondence theory. (See §1.4.2 of Field 1972 for the implementation of this theory that is generally considered to be a correspondence view, and see Davidson 1977 for the opposing position.) Much of the remainder of this chapter, along with chapter two, will be devoted to this issue.

[10] And perhaps *properties* for the predicates to express, but this is an issue that we are bracketing.

truth-bearers were changed when we switched from the classical correspondence theory to the modern correspondence theory. For the classical theory, it is typically assumed that truth is primarily a property of propositions. But when presenting the modern object-based theory, we concentrated on *sentences*—i.e. language-bound vehicles of communication. So why the change?

The main reason to focus on sentences (at first) is for ease of exposition. One of the essential claims of the modern correspondence theory is that truth conditions are explained by compositional rules that are sensitive to syntax. Well, compared to propositions, we have a relatively good grasp of the syntax of sentences. This makes sentences the preferred object of study within the context of this theory. Propositions, by contrast, are comparatively obscure. (Of course, there are philosophical theories that purport to describe the structure of propositions, but there is no such theory that commands universal assent.) It is better to develop the theory where we can—that is, the domain of linguistic truth-bearers—before moving on to relatively less well-understood domains.

The object-based correspondence theory does not necessarily shun the existence of propositions. In developing the theory, it may turn out that introducing these entities is inevitable and that they exhibit the right compositional structure to work for this theory of truth. The theory does not rule this possibility out; it can, at first, maintain a degree of agnosticism about the domain of truth-bearers (Glanzberg 2015). We will return to the issue of truth-bearers in the penultimate section of this chapter, but I make no pretence to cover this topic in all of its depth here.

1.5 Semantic composition

For our running example thus far, we have focused on a straightforward sentence of subject-predicate form ('this is gold'). Since the syntactic structure of this sentence is so simple, it was obvious how to state the rule for semantic composition: the sentence is true if, and only if, the predicate term applies to the thing referred to by the subject term. Now, despite the seeming obviousness of this rule, this *is* a significant step forward from the fact-based theories we considered earlier. This is because, unlike the fact-based theories, we now have a general explanation of the truth conditions for all subject-predicate sentences, provided an account of the primitive semantic facts.

Of course, this doesn't yet say anything about the truth conditions for other syntactic forms. And in fact, one reason that the object-based theory is not so straightforward to articulate in full generality is precisely that most sentences aren't this simple. Within any natural language that humans actually speak, the variety of available sentence structures is enormous. Hence, to flesh out the theory and elevate it above the 'toy' version, we must provide more rules of semantic composition to cover the variety of available sentence structures. And to maintain the explanatory integrity of the theory, we should strive to do this fairly systematically. To this end, we should borrow an insight from logical atomism: ideally, the rules should be *recursive* so then we can explain a potential infinity of truth conditions from a finite stock of rules.[11]

There are two fields of inquiry that have contributed to filling in this gap. Specifically, this approach to explaining truth depends on contributions from logic and linguistics to provide its theory of semantic composition. In the interest of surveying this conception of truth, I will need to say a bit about both of them.

1.5.1 Tarski's theory of truth

It is not possible to discuss the topic of truth for very long without mentioning the Polish logician, Alfred Tarski. Tarski's widely-celebrated method for defining truth is significant for several reasons (see his 1944). For one, it provides a complete, explicit, *recursive* theory of semantic composition for a central class of artificial formal languages. For another, Tarski himself self-consciously expressed it *as a theory of truth.* But his approach also comes with a crucial limitation: according to the popular imagination, it is only designed to apply to certain *artificial* formal languages.[12]

To illustrate Tarski's method, I will follow the common practice of focussing on the formal languages defined for the systems of first-order logic. To construct such a language, we must delineate a system of variables ('x_1', 'x_2', 'x_3', …), names ('c_1', 'c_2', 'c_3', …), predicate symbols ('p_1', 'p_2', 'p_3', …), the first-order quantifiers ('$\forall$' and '$\exists$') and the propositional connectives ('$\neg$', '$\rightarrow$', '$\vee$', and '$\wedge$'). We then define the syntactically permissible sentences and formulae (the WFFs) using standard recursive methods.

It is important to stress that these languages are wholly artificial. This means that they are *intelligently designed* by an artificer, and as such, the artificer has the license to stipulate each of the language's pertinent features. When we define the class of WFFs, for example, we may do so without consulting any empirical theory of syntax. Likewise, when it comes to the primitive

[11] This calls for another historical comment. The project of providing a compositional semantic theory that matches the full variety of sentential syntactic structures is really a project of the 20th and 21st centuries. The logical apparatus required to enact this project was only developed as recently as Gottlob Frege in his *Begriffsschrift* (1879). Before then, during the early modern period in Europe, it was common for philosophers to assimilate all 'judgments' to the subject-predicate form (this was owed to the influence of Aristotle's syllogistic logic). This is one reason why it is apt to call the present correspondence theory, which avails itself of these developments, a 'modern' correspondence theory.

[12] The historical Tarski was in fact interested in a broader range of languages and applications than discussed here. But to simplify the exposition, I will follow the now-common practice of construing the theory as primarily concerned with artificial first-order languages.

semantic properties of the language, we may stipulate their semantic contributions by fiat. We can define '∨' to mean disjunction, '¬' to mean negation, 'c_1' to refer to one thing and 'c_2' to another; and once these semantic properties have been declared, there can be no further mystery as to how they were determined.

For these reasons, it follows that these languages are highly controlled objects to study. When such a language is the object of investigation, we call it the *object language.* This distinguishes it from the *metalanguage* that we use to talk *about* the object language. It is necessary to distinguish these two languages, especially when discussing Tarski's theory. That is because, according to Tarski, an object language *L* must not be allowed to contain its own truth predicate. Instead, the truth predicate for the sentences of *L* will be an expression of the metalanguage (1944: 380–1). This means that we cannot formulate within the object language *L* any sentence that asserts the truth or falsity of any sentence of *L*, so we sidestep the troubles of the liar's paradox.

Since we have so much sovereignty over these formal languages, they are ideal subjects for explicit definitions of truth. They are particularly congenial to recursive definitions that follow the syntax since the syntax is explicitly laid out in the definition of the language. It is worth specifying Tarski's account since I will refer back to it often.

The steps to defining *true-in-L* for an object language *L* first involve defining reference and application for the simple names and predicates. We pick a domain D = {s_1, s_2, s_3, ...} of objects which will serve as our domain of discourse for *L*. Then, since our language is artificial, we can stipulate what each name refers to by providing a list: i.e. *'c_1' refers to s_1, 'c_2' refers to s_2*, ... for each name in the language. Likewise, we can stipulate the application of each predicate: *'p_1' applies to s_j, s_k*, ... for each thing that 'p_1' will apply to, *'p_2' applies to s_m, s_n*, ..., and so on.[13] Each of these lists provides our base clauses for the definition.

Once we have the base clauses, we can then proceed to define application and truth for the rest of the formulae of *L*.[14] First, we define *denotes$_S$ (denotation relative to the sequence S = <s_1, s_2, s_3,...> of objects from D)* as follows:

1. '$\underline{x_k}$' denotes$_s$ s_k (for *k* = *1*, *2*, *3*, …)
2. '$\underline{c_k}$' denotes$_s$ what it refers to.

Then we recursively define *satisfaction$_S$* (*satisfaction by the sequence S*):

[13] 's_j', 's_k', 's_m', 's_n' are dummy names to represent the objects applied to by the predicates.

[14] P is a schematic letter representing arbitrary predicates from *L*; t represents the terms of *L;* Φ and Ψ represent the formulas of *L;* and the underlined '$\underline{\neg}$', '$\underline{\wedge}$', '$\underline{\vee}$', '$\underline{\rightarrow}$', '$\underline{\forall}$', '$\underline{\exists}$', '$\underline{(}$', and '$\underline{)}$' are metalinguistic representatives for the object language terms '¬', '∧', '∨', '→', '∀', '∃', '(', and ')' respectively. '$\underline{x_k}$' and '$\underline{c_k}$' are metalinguistic representatives for the object-language variables and names.

3. An atomic formula $\ulcorner P(t) \urcorner$ consisting of a predicate P and a term t (a name or variable) is satisfied$_s$ if and only if

(i) there is an object *o* that t denotes$_s$
and (ii) P applies to *o.*

4. $\ulcorner \underline{\neg}\ \Phi \urcorner$ is satisfied$_s$ if and only if Φ is not satisfied$_s$.
5. $\ulcorner \Phi \ \underline{\wedge}\ \Psi \urcorner$ is satisfied$_s$ if and only if Φ is satisfied$_s$ and so is Ψ.
6. $\ulcorner \Phi \ \underline{\vee}\ \Psi \urcorner$ is satisfied$_s$ if and only if Φ is satisfied$_s$ or Ψ is satisfied$_s$.
7. $\ulcorner \Phi \ \underline{\rightarrow}\ \Psi \urcorner$ is satisfied$_s$ if and only if either Φ is not satisfied$_s$ or Ψ is satified$_s$.
8. $\ulcorner \underline{\forall x_k}\ \Phi \urcorner$ is satisfied$_s$ if and only if for each sequence S* that differs from S at the *k*th place at most, Φ is satisfied$_{s*}$.
9. $\ulcorner \underline{\exists x_k}\ \Phi \urcorner$ is satisfied$_s$ if and only if there is a sequence *S** that differs from *S* at the *k*th place at most and Φ is satisfied$_{s*}$.

Finally, we define a sentence as *true-in-L* if and only if it is *satisfieds* for all *S*. Although this definition is recursive, Tarski proved that it can be converted into an explicit definition of the form: *for all expressions e in L, e is true if and only if F(e)* (where F is an explicit definition of truth).

The key takeaway is that Tarski achieved, for first-order formal languages, exactly what an object-based correspondence theory is supposed to do. His theory effectively generates the truth conditions for any arbitrary sentence of any such object language based on its syntactic structure, a finite set of recursive rules, and the semantic properties of its primitive parts. The improvement over our toy theory from earlier can hardly be understated. We can now account for the truth conditions of *any* sentence of a first-order language, no matter the complexity.

Nonetheless, we must also be clear on the limits and scope of Tarski's accomplishment. To repeat, Tarski only applied his theory to the sentences of artificial formal languages. He expressly declined to apply his theory to the sentences of natural language, which he regarded as intractable (1944: 376). He thus gave himself a highly circumscribed target.

Subsequent debate in the philosophical literature over the significance of Tarski's definition has centred mainly around the extent to which we can appropriate the Tarskian methods for natural language. (This is to say nothing of its undisputed significance in other technical areas besides philosophy—e.g. logic, mathematics, and computer science.) There are several dimensions to this question, but the one that matters most to present purposes is the extent to which Tarskian methods can be extended toward a general account of *truth.* This is the topic which we will turn to next.

1.5.2 Field on Tarski

The next development for the correspondence theory comes from ‘Tarski’s Theory of Truth’ (1972) by Hartry Field. In this paper, Field takes several steps towards developing a theory that is based on Tarskian principles. This paper is especially significant for us because it is one of the earliest pieces that provide the blueprints for what I’m calling the modern object-based correspondence theory.

Even though both Tarski and Field are nominally interested in characterizing truth, it must be stated upfront that Field had different ambitions than Tarski. Tarski’s aim was to show that truth is a scientifically respectable notion, and so he limited the scope of his theory to certain formalized languages. Field, on the other hand, was interested in characterizing truth for the whole range of natural languages (1972: 348). This gives Field’s envisioned theory a claim to generality that isn’t present in Tarski’s theory, but it also means that he cannot straightforwardly adopt Tarski’s original definition. Since he had this objective, the first task for Field’s paper is to isolate the elements of Tarski’s theory that are serviceable to his own aims.

Field’s first major claim is that a Tarskian definition of truth for a *natural* language (a language that humans actually speak) would be inadequate for the philosophical purposes for which we might seek a definition of truth. Field is quite explicit as to what those aims are. At the time of his writing, the prevailing opinion was that Tarski’s theory succeeded in reconciling semantics with the broadly scientific worldview (1972: 347, 359). Tarski’s theory allegedly achieved this, according to the received opinion, because it showed how to define *truth* and *reference* in non-semantic terms, rendering these notions innocent from the point of view of the fundamental sciences (356–7). However, as Field argues, Tarski did not achieve any such thing regarding the *primitive* semantic facts for *natural* languages.

The issue revolves around the base clauses in a Tarskian truth definition. When we define reference in the context of a Tarskian definition for a language *L*, we essentially do so by stipulating a list of name-referent pairs. We may define refers-in-*L* explicitly with a definition of the form: *for all x, for all y, x refers to y if, and only if* $x =$ ‘c_1’ *and* $y = s_1$, *or* $x =$ ‘c_2’ *and* $y = s_2$, *or* ... (and so on). Likewise, we may define application (or reference) for predicates with a definition of the form: *for all x, for all y, x applies to y if, and only if,* $x =$ ‘p_1’ *and* $y = s_j$, *or* $x =$ ‘p_1’ *and* $y = s_k$, *or…* (including a clause for each thing ‘p_1’ applies to), *or* $x =$ ‘p_2’ *and* $y = s_m$, *or* $x =$ ‘p_2’ *and* $y = s_n$, *or* (including a clause for each thing ‘p_2’ applies to), *or* ... (and so on). As I have emphasized earlier, this method for characterizing the primitive semantic facts is entirely appropriate *if the object language is artificial.* That is because the semantic facts for artificial languages are determined by the artificer’s fiat.

But if, instead, the language in question is natural, then, according to Field, a list-like definition of reference would be inadequate to capture the phenomena. Reference for a spoken language must be determined somehow by the usage of its speakers, but a Tarskian theory cannot capture how this works. For this reason, Field argues that we cannot rely on a purely Tarskian theory to account for truth in natural languages. We also need an additional account of how the facts of usage for a natural language determine the primitive semantic facts for that language.[15]

Nonetheless, Field sees Tarski's theory as composed of two detachable parts. It comprises the list-like accounts of the primitive semantic facts (which appear in the base clauses) and the recursive clauses that define satisfaction and truth for semantically complex expressions. So even if the former would be inadequate to serve a general account of truth, Tarski's theory is still not without value. According to Field, the sole philosophical significance of Tarski's theory resides squarely in the latter component of the theory (370). Regardless of the base clauses, the recursive clauses still generate the truth conditions of each sentence of a first-order language based on the reference properties of the simple names and predicates of the language. So then, if we could append the base clauses with an appropriate account of primitive reference, we could utilize the rest of the Tarskian apparatus to provide a substantive account of truth (for first-order languages). In Field's opinion at the time, Tarski's chief philosophical achievement was that he reduced the problem of explaining truth to the problem of explaining reference (347).

All of this suggests a two-pronged approach to developing a theory of truth. Towards the end of his paper, Field lays out his agenda for the remaining tasks at hand. The theory he envisions exhibits the same division of labour as the modern object-based correspondence theory described in §1.4. It contains one component that recursively explains sentential truth conditions based on the semantic contributions of simpler expressions. This is provided by Tarski's definition minus the list-like accounts of primitive reference. It then contains another component to account for reference for the primitive expressions. Once this is achieved, Field suggests that the two may be synthesized into an overall account of truth.

Since Field's envisioned theory incorporates this division of labour, he argues that the entire account of truth was not yet complete. Tarski may have achieved the first step, but the final frontier was to uncover a theory of reference for the primitive expressions of natural language.[16] He thus calls for a programmatic effort to develop such a theory.

Throughout the paper, Field makes several suggestions as to what more is needed to account for the primitive semantic facts. For one, he says that the account should analyze the

[15] Field argues for this conclusion with an analogy. He imagines a scenario where early chemists propose an account of chemical valence by merely listing off the pairs of elements and valences. He then argues that this procedure alone would be insufficient for reducing the chemical property of *valence* to physical properties according to the standards of proper scientific methodology because it doesn't explain *how* the subatomic structures determine the valence properties. Likewise, Field claims that a list of word-referent pairs would fail to reduce reference (according to the rigours of proper science) because it fails to adequately explain the pairings (362–5).

[16] Field writes that the reductionist aim "rules out the possibility of [Tarski's definition] *by itself* being an adequate truth definition; and it is right to do so, if the task of a truth definition is to reduce truth to non-semantic terms, for [the recursive clauses] provide only a *partial* reduction. (To complete the reduction we need to reduce primitive denotation to non semantic terms.)" (362)

notion of reference in *non-semantic* terms (e.g. 360). Hence it must not mention *satisfaction, truth, meaning,* etc. In addition, Field often speaks of the need to 'reduce' reference to non-semantic relations (e.g. ibid). However, the crux of his argument is that we cannot achieve this reductionist constraint by merely listing off word-referent pairs (*à la* Tarski). This suggests that the account must be, to some degree, *unifying;* as much as possible, it must strive to say what the instances of reference *have in common.* Taken together, this theory of truth requires an account of reference that has the form:

(IR) For all x, for all y, x refers to y iff x bears relation R to y,

where R is specified in non-semantic terms. (I call this form 'IR' for *inflationist reference* since it is one of the high-water marks of the inflationary conception of truth and reference. This contrasts with *deflationary* approaches, which will be explained in chapter two.) Field proposes, at the time of this article, that this program might be carried out by expanding the newly-developed causal theories of Kripke and Putnam (367; see Kripke 1980, Putnam 1973, 1975). He also suggests that the mechanisms of reference are likely to be uncovered by investigations into psychology and neurophysiology (373).

It is fair to ask why Field assumes that the theory of reference must take this form. Why must the theory be cast in *non-semantic* terms? For Field, the reason has all to do with the doctrine of physicalism. As he explains, physicalism is a methodological commitment to accept into one's ontology only those objects, properties, or relations that are explicable in physical terms. So reference must be explicable in terms of *non*-semantic relations because that is the only hope for making truth and reference acceptable in light of this commitment. He also writes that "if… we were to ever conclude that it was *impossible* to explicate the notion of truth and [reference] in nonsemantic terms, we would have either to give up these semantic terms or else to reject physicalism" (1972: 360).

We will discuss several of the attempts to realize Field's vision for the theory of reference in section §1.6. But for now, it is worth pausing to observe another difficulty for this general theory of truth. Field takes it for granted that it is possible to apply Tarski's recursive clauses to yield the compositional component of the theory *for the sentences of natural language.* But Tarski's theory (as I've been construing it) was specifically limited to *formalized, first-order languages.* Since these languages are formal, they have relatively straightforward syntactic structures, which allow for relatively straightforward characterizations of the semantic rules that are sensitive to this structure. And since these languages are first-order, they leave out many of the kinds of complicated sentence structures found in natural language.

There are really two kinds of problems here. On the one hand, many varieties of natural language sentences are prima facie truth-evaluable and yet cannot be translated into any equivalent sentence with first-order syntax. This includes modal sentences, sentences with higher-order or plural quantifiers, probabilistic sentences, sentences with indicative conditionals, and generics. On the other hand, even when we *can* translate a sentence into the language of quantificational logic, it is still often not the case that the original sentence will have the same syntactic structure as its formal translation. For instance, 'all dogs bark' will be translated as '$\forall x(Dx \rightarrow Bx)$', but clearly, the former is not syntactically equivalent to the latter. However, Tarski's theory only provides compositional rules that are sensitive to the syntax of the latter. So if Field wants to claim that Tarski's theory can account for the truth of the former, then he needs some story to tell to make up for this difference.

(There is a third difficulty for Field's theory that I do not have the space to delve into here. By broadening the scope to include natural languages, Field is thus aiming to characterize truth for languages that contain their own truth predicates. He is thus closing himself off from Tarski's solution to the paradoxes of self-reference. Field would therefore need some other means to handle these paradoxes to carry out his project. Later in his career, after changing his mind and preferring a different conception of truth—see chapter two—Field offered his own solution (2008). Unfortunately, I do not have the space to discuss the paradoxes of self-reference.)

1.5.3 Expanded logics and the Fregean program

The main challenge to Field's theory just described amounts to this. Field had assumed that Tarski's recursive clauses could provide the compositional component of an object-based correspondence theory of truth. But once we switch from considering formal, first-order languages to natural language, the Tarskian theory alone is no longer comprehensive enough to handle the full variety of sentence structures.[17] A more powerful theory of semantic composition is needed to deliver the truth conditions for the wider variety of natural language sentences.

To answer this challenge, we would have to attend to the enormous body of work done on semantics in both logic and linguistics. However, the purpose of this chapter is not to provide a detailed survey of formal semantic theory. It is impossible to survey all of the developments on this front since Field's writing in 1972. Nonetheless, an object-based correspondence theory needs *some* response to the above concern. So, to this end, a few rapid remarks will have to suffice.

The first remark is that the object-based correspondence theory can avail itself of any of the numerous expansions on first-order logic if it suits its purposes. The semantic theories for modal logic, higher-order logic, temporal logic, plural quantifiers, counterfactuals, etc. all fulfill

[17] *Pace* the semantic program of Davidson (1967).

the needed task for the composition component of a correspondence theory: they provide rules for generating the satisfaction and truth conditions for complex formulas and sentences on the basis of their simpler parts.[18] Each of these can be viewed as a formal representation of how the expanded vocabulary (modals, tense, higher-order quantifiers, non-classical conditionals) contribute to the truth conditions and entailment relations of the sentences that are their hosts.

With that said, it must be mentioned that the semantic theories of the expanded logics carry novel questions of interpretation that aren't present in first-order semantic theory. Most notoriously, the semantic approach to modal logic quantifies over points of evaluation that are most naturally interpreted as possible worlds. Within the context of a *correspondence* theory of truth, this raises distinctly philosophical questions about ontological commitment. To wit, must the correspondence theorist accept the existence of these exotic metaphysical postulates in order to press into service the semantic theories of the expanded logics?

Much has been written concerning the controversy surrounding the apparent ontological commitments of the expanded and higher-order logical theories.[19] Once again, this is not a battleground that I wish to enter for present purposes (although I will offer some conciliatory remarks at the end of this section). The chief reason for mentioning these controversies is that this is an issue on which the correspondence theorist cannot remain forever neutral.

The second line of inquiry that the correspondence theorist may wish to appropriate is the Fregean semantic theory produced in the generative tradition.[20] Here, the basic assumption is that the generative theory of syntax provides the 'real' syntactic structure of natural language. Following this, the semanticist endeavours to account for semantic composition by utilizing the Fregean assumption that lexical meanings compose by functional application. To this end, they assign denotations to expressions that are either objects, truth values, or some function defined in terms of objects and truth values.

The effort to integrate a Fregean semantic theory with generative syntax is both successful and ongoing. It also rectifies one of the disadvantages of applying the Tarskian theory to natural language: we can take the syntax of natural language realistically without artificially shoehorning it into the syntax of quantificational logic. But for the correspondence theorist, it bears mentioning that Fregean semantics is also not without its own questions of philosophical interpretation.

This time, the puzzle stems from the fact that a Fregean semantic theory will assign complex *mathematical objects* (namely, functions) as the denotations of most significant expressions (all except for sentences and proper names). Once again, this raises questions of

[18] Sider (2010) provides an overview.

[19] See e.g. Lewis (1986).

[20] See Heim & Kratzer (1998).

realism within the context of a correspondence theory. The correspondence theory claims that truth is built upon subsentential semantic facts and a Fregean theory assigns mathematical objects as subsentential denotations. One must wonder whether truth is really a matter of representing these *mathematical* objects and their interrelations.

Consider, for example, the sentence ‘all dogs bark’. Intuitively, this is true in virtue of representing all of the dogs as barkers, which, let’s suppose, they all are. Pre-theoretically, we would say that the sentence is about *dogs* and *barking.* But according to a Fregean semantic theory, the denotations of ‘barks’ is a function from objects to truth-values (which outputs true to any barker, and outputs false otherwise). Worse still, the denotation of ‘all dogs’ is a higher-order function that inputs functions (from objects to truth-values) and outputs a truth-value (the denotation of ‘all dogs’ will return true if the input function is such that it outputs true for all dogs). This time, the problem isn’t just a matter of ontological commitment to mathematical entities. Rather, the problem concerns how we reconcile our pre-theoretical judgments about the subject matter with the exotic posits of a Fregean semantic theory (see Simchen 2017: ch. 3).

So, to recap the present situation, the object-based correspondence theory requires a theory of semantic composition beyond Tarski’s first-order theory. And to this end, there are plenty of *formalisms* in the offing, which vary depending on their theoretical aims (e.g. whether they subject themselves to the constraints imposed by generative syntax). However, these formalisms introduce entities and denotations which are surprising from the point of view of our naive judgments of ontology and subject matter. The question here is how to conceive of the compositional component of the correspondence theory in light of these pronouncements from contemporary semantic theory.

In this author’s opinion, it is appropriate to understand the formalisms that *represent* semantic composition with a modicum of instrumentalism. We need not automatically read into the semantics of modal logic a realist thesis about possible worlds, and we need not automatically read into Fregean semantic theory any claim to the effect that ordinary speakers are talking about set-theoretic objects. In each case, the introduction of these entities serves important modelling purposes within the specific aims and constraints of the formalism, but appreciating this point does not demand that we revise our opinions on ontology or subject matter for the sake of the correspondence theory of truth. To fully argue this point would take me too far afield (see Simchen 2017: ch. 3).

The fact that these formal semantic theories have been developed with considerable richness and flexibility should allay any worries about whether the correspondence theory can deliver on its compositional component.[21] Hence, going forward, I will assume that some such component is available. That is, I will assume that the truth condition of each sentence is

[21] See Pagin & Westerståhl (2010a) and Pagin & Westerståhl (2010b) for the present state of this question.

(somehow) determined by the semantic facts pertaining to its subsentential components by semantic means. How exactly this determination occurs will not be my primary concern.

1.6 The theory of reference

A workable correspondence theory of truth needs an account of the 'correspondence' relation between the truth-bearers and the portions of reality to which they correspond. As per §1.4, a modern object-based correspondence theory accounts for sentential correspondence to reality (i.e. truth conditions) by decomposing it into subsentential correspondence to objects (reference relations). The approach can thus be understood as the product of two factors: a theory of semantic composition and a theory of primitive semantic facts. Whereas we have seen in §1.5 that the former is largely within the domain of logic and linguistics, the second component—the theory of reference—is much more at home in philosophy. It is the second component to which we will now turn.

Broadly construed, the theory of reference concerns the relations between a language's simple lexical (non-logical) expressions and the worldly objects they represent. In other words, it is the theory of what the basic symbols symbolize, and why they come to represent what they symbolize. It is concerned with *word-world relations.*

The purpose of this section is two-fold. First, I would like to distinguish and classify the various questions or explananda that the traditional theories of reference intended to answer. These distinctions will be necessary for subsequent chapters. Secondly, I will quickly survey the prominent contributions to this line of inquiry.

1.6.1 Distinctions

The most important distinction to make upfront is between semantics and metasemantics. This terminology comes from Kaplan (1989a):

> There are several interesting issues concerning what belongs to semantics. The fact that a word or phrase *has* a certain meaning clearly belongs to semantics. On the other hand, a claim about the *basis* for ascribing a certain meaning to a word or phrase does not belong to semantics. 'Ohsnay' means *snow* in Pig-Latin. That's a semantic fact about Pig-Latin. The *reason* why 'ohsnay' means *snow* is not a semantic fact; it is some kind of historical or sociological fact about Pig-Latin. Perhaps, because it relates to how the language is

> *used*, it should be categorized as part of the *pragmatics* of Pig-Latin… or perhaps, because it is a fact *about* semantics, as part of the *Metasemantics* of Pig-Latin. (573–4)

The fundamental distinction here is between *what* a given expression *means* or *refers to* and *why* it means (refers to) what it does. (In the context of a truth-conditional semantic theory, meaning and referring are closely related; for singular terms, they amount to the same thing.) *Semantics* is the level at which we are concerned with identifying the referent of a term or applications of a predicate, and *metasemantics* is the level at which we are concerned with *explaining* these semantic facts. In Simchen (2017), the elementary task of metasemantics is rendered into the question, "What determines that expressions have their semantic significance?" (2). Burgess and Sherman add that the metasemantic grounds for semantic facts should be *non-semantic* (Burgess & Sherman 2014). So a metasemantic explanation for a semantic fact is ultimately one that uncovers the underlying non-semantic facts that determine the semantic fact. It is important to emphasize that this pursuit is conspicuously a matter of metaphysics; it is asking what *metaphysically determines* or *grounds* the facts of semantics. It is not a matter of local epistemology; it is not asking how we come to *know* the facts of semantics.[22]

Although semantics and metasemantics are complementary projects, there is also a sense in which they operate independently. A complete semantic theory for a language, which assigns semantic values to every significant expression, must get the semantic facts right for the primitive terms of the language. However, for its own specific aims, the theory need not be sensitive to *how* these semantic facts are determined by underlying, non-semantic states (Dickie 2015: 11). On the other hand, a metasemantic theory must also get the semantic facts right for the primitive terms of the language. But for its aims, it need not be sensitive to the modelling techniques used by the various semantic theories for their own intratheoretic purposes (Simchen 2017: 82–89). Semantics and metasemantics thus meet at a point (agreement over the basic semantic facts) and they inform each other in subtle ways. But since their explanatory purposes differ, they each enjoy some degree of relative autonomy.

It is clear that an object-based correspondence theory must incorporate the semantic facts for each referring term. The theory needs an assignment of referents to deliver the truth conditions of all of the truth-bearing sentences. However, it is also necessary for a correspondence theory that it can answer metasemantic questions as well. Semantics alone will tell us that '*this*' (when used in the context of our example) refers to the rock in my hand, that 'is gold' applies to gold things, and so *ipso facto* that *'this is gold' is true if and only if the rock in my hand is gold.* But if our overall explanatory endeavour is to 'explain the nature of truth', then,

[22] In the original quotation from Kaplan, he assumes that the metasemantic facts are 'historical' or 'sociological'. Later writers have narrowed the scope of metasemantics so that it pertains, specifically, to questions of metaphysical grounding. Moreover, there may be other grounds of semantics besides history and sociology; there's also cognitive science (for example).

presumably, we are also indebted with explaining *why* the correspondence relations obtain on the subsentential level. It is not enough to simply be told that my token ‘*that*’ refers to the rock in my hand, and that ‘snow’ refers to snow, and ‘grass’ refers to grass, and so on—as we list off the referents of each term. In addition to identifying all of the semantic facts, we also need to explain them. This is the kernel of truth that Field uncovered in his (1972): a truth-conditional semantic theory (like Tarski’s) is key to a general metaphysics of truth, but to satisfy the philosopher’s aims, the theory must also be accompanied by a metasemantic account. (Granted, the historical Field from (1972) did not put the point this way. He was then concerned with reduction, not grounding. This is a discrepancy to which we will return.)

So the modern object-based correspondence theory that we have been exploring requires a theory of metasemantics. Not only does it need to supply the facts of reference, but it also needs to explain them. The next step is to look at the options for how this might be done.

Given the overall architecture of this correspondence theory, the theory is already limited in the kinds of metasemantic accounts that it may find acceptable. The reason has to do with the order of explanation prescribed by the theory. An object-based correspondence theory seeks to explain truth based on the representational properties of subsentential expressions. It thereby requires that *reference* come before *truth* in the order of metaphysical explanation. Truth is (partly) *grounded* in reference on this view. By giving truth and reference this relative explanatory ordering, it then follows that the correspondence theorist must seek grounds of reference that do not depend on truth. This means that they mustn’t look to ‘holistic’ theories of reference determination. Instead, their metasemantics must be *atomistic* (in the sense of Fodor 1998): it must explain reference determination for each term individually, without presupposing the truth conditions of bodies of sentences.

There is a fundamental choice point within metasemantics between *productivist* and *interpretationist* approaches (Simchen 2017). To a first approximation, a productivist seeks to ground the facts of reference for each referring expression within the factors surrounding its production. To explain the reference of any given expression, a productivist may, for example, appeal to the referential intentions of its speaker (*à la* Donnellan 1966), the etiology of the term (*à la* Putnam 1975, Kripke 1980), or the referential function conferred upon the term (*à la* Millikan 1984).

This is opposed to the interpretationist, who instead seeks to explain the determination of reference in terms of the *interpretive consumption* of the given referring expression (Simchen 2017: 4). According to the interpretationist, what it is for an expression to have a certain meaning —or for a term to refer to a particular thing—*just is* for the expression or term to be interpretable as such. Moreover, the interpretation of any single expression takes place within an interpretive theory of the subject’s overall linguistic behaviour. The approach is therefore holistic. One example is Davidson’s view that meaning is constituted by a Tarskian truth definition that can explain and predict the subject's overall verbal behaviour and rationality (Davidson 1967, 1974, 1977). Another example is Lewis' view that a semantics for a language (what he calls a grammar) is determined by what best explains the overall assignment of truth conditions to the language's sentences, which are, in turn, determined by conventions of truthfulness and trust in that language among its speakers (Lewis 1969, 1974, 2013). (See chapter four for more details.)

Not only is interpretationism holistic, but it also places truth before reference in the order of explanation. In each fleshed-out version of interpretationism, the assignment of referents to names is determined by a prior assignment of truth conditions to sentences. For this reason, we cannot combine an object-based correspondence theory of truth with an interpretationist orientation to metasemantics. For our running conception of truth to work, we must adopt the productivist framework.

Before we examine the various theories that fall under the productivist heading, there is one final distinction to make. As we defined it, metasemantics is primarily concerned with *grounding*; it is concerned to identify the non-semantic *grounds* of the semantic facts. Otherwise put, it is concerned with the *determination* or *fixation* of the facts of reference. For instance, we may be interested with explaining why my token of '*this*' refers to the gold in my hand rather than any other object. Such an endeavour would paradigmatically fall within the domain of metasemantics.

Now, although the traditional theories of reference were interested in solving the grounding problem, they were also keen on another task. (They also tended to run these two tasks together.) Specifically, the traditional theories were concerned with *explaining what reference itself is.* They had the aspiration to uncover the *nature* of the reference relation. (Just as chemists discovered that water is H_20, one might hope that linguists or cognitive scientists would uncover a relation that's identical to reference.) This means that ultimately they were looking to fill in the details of Field's template—that is, an account of the form:

(IR) For all x, for all y, x refers to y iff x bears relation R to y.

Doing so would then allow one to say that *'a's referring to b* just is *'a's bearing relation R to b.*

We should appreciate that the task of explaining reference fixation and analyzing the nature of reference are two distinct projects. It is true that both of them fall under the broad heading of 'metasemantics' since each aims to explain some dimension of the semantic facts. It is also true that an *analysis of reference* might illuminate the issue of what determines reference (if such an analysis were to be forthcoming). But we should not begin by conflating the two. Each of them targets a different explanandum. Take a routine example of a semantic fact: that *'a' refers to b*. One central task for metasemantics is to focus on the referent, *b,* and ask *why is it b rather than c or d that was determined as the referent.* In contrast, the latter project focuses on the '*refers*' part and asks *what this relation consists in*.[23] In subsequent chapters, I will call the former a *selective explanation* of reference since the point is to explain why a particular object was *selected as the referent,* and I will call the latter *an account of reference's nature.*

(Selective explanations form a core component of the purview of metasemantics. But they are not the only part. Another task for metasemantics is to explain why it is that a given referring item is endowed with intentionality. In other words, metasemantics also seeks to explain how reference relations are generated.)

1.6.2 Theories

Let's now pick up the thread where we last left it with Field (1972). Recall that Field argues that the correspondence theory needs an IR account (i.e. an *analysis*) of the reference relation. He also gestures towards the causal theories to fill in this gap.

Field was writing around the same time that Saul Kripke and Hilary Putnam were revolutionizing the way philosophers thought about reference. Before then, it was widely assumed that a term or a concept used by a subject would refer in virtue of being associated with a set of descriptive conditions within the subject's cognitive grasp, which would (in turn) be satisfied by the referent (see e.g. Russell 1919[24]). But by Kripke and Putnam's joint efforts, this consensus was overturned. This paved the way for a new way of thinking about reference, whereby reference is determined by factors that are external to the subject.

Although the resulting pictures are similar, Kripke and Putnam each have different motivations and emphases in their theories. Putnam, for instance, was particularly keen on explaining how distinct stages of scientific theory can share the same subject matter, despite making vastly different claims about their subject matter (1973, 1975). This is important for him because it allows one to see science as *progressing* in the sense of improving our knowledge of a common subject matter.[25] Because of his ulterior motives in the philosophy of science, he concerns himself primarily with natural substances and kinds—e.g. water, tigers, and gold. The concern was to explain how, for instance, we can be referring to the same stuff as the ancient Greeks when we use our term for water and they use theirs. After all, we say that water is a compound and they say that water is an element. So how can the sameness of subject matter be secured?

[23] Burgess & Sherman (2014) draw the same distinction and call the former task *basic metasemantics* and the latter task *the theory of meaning.*

[24] And see Donnellan (1966) for an earlier instance of dissent from the descriptivist paradigm.

[25] This is opposed to the incommensurability thesis of Feyerabend (1962) and Kuhn (1962).

To put it crudely, Putnam's answer is that we mustn't look to the descriptions or theories offered by the speakers to do any reference-fixing work. Instead, we should appeal to the fact that each scientific generation is causally interacting with the same natural environment. According to his causal theory, the referent of a given natural kind term (e.g. 'water') is determined by its causal history—in particular, by what the scientists actually demonstrate when they introduce the term (1975: 149).[26] As long as we can trace back the history of the various terms to demonstrations of the same kind of substance, the theories surrounding each term will share the same subject matter.

In light of the contrast from the previous subsection between selective explanations of reference and analyses of reference, it is worth pointing out that Putnam's thesis can sensibly be understood as a species of selective explanation. The point, for Putnam, is to explain how two terms can have the same referent *rather than distinct referents.* But to explain the sameness of reference does not require an analysis of what reference consists in; it only requires an explanation of what determines the referent of each term.

Whereas Putnam's theory was motivated by a scientific realist agenda, Kripke had relatively less interest in the philosophy of science. The focus of much of his (1980) discussion concerns the proper names for individuals as they occur in everyday discourse. He is interested in how average speakers refer with their use of ordinary names, such as 'London' and 'Aristotle'. To give a characteristic example, one of Kripke's many polemics against the descriptivist paradigm is that the ordinary speaker who uses the name 'Aristotle' need not command any description that can single out Aristotle himself (1980: 81).

After he deposes the descriptivist accounts, Kripke offers his own positive picture, which holds that the primary mechanism for reference for ordinary speakers is social deference (91). When the average person uses the name 'Aristotle', they may not have any description in mind that singles out the referent. Instead, they intend to use the term to refer to the same person (object) as was referred to by whomever they learned the term from. When speakers pass a term from one to another in this way, they form a chain that traces the term's history of usage. Since the chain for 'Aristotle' traces back to Aristotle himself (through his associates who gave him the name in an initial 'baptism'), this socio-historical fact determines that 'Aristotle' refers to Aristotle.

Kripke is explicit that he only intends this positive account to offer a 'picture' of reference fixation and not to constitute a *theory* of reference. In one famous passage, he writes,

> One might never reach a set of necessary and sufficient conditions [for reference]. I don't know, I'm always sympathetic to Bishop Butler's 'Everything is what it is and not

[26] Putnam's explicit condition is given by: "(2') (For every world W) (For every x in W) (x is water = x bears same$_L$ to the entity referred to as 'this' in the actual *world* W1)" (1975: 149), where 'same$_L$' denotes sameness of chemical kind.

> another thing'—in the nontrivial sense that philosophical analyses of some concept like reference, in completely different terms which make no mention of reference, are very apt to fail. (94)

To put this into our terminology, Kripke's causal account only aims to tell us the mechanisms that are relevant to selective explanations (for ordinary proper names), and he openly declines to offer an account of reference itself.

Putnam and Kripke's contributions are promising to those, like Field, who want a full-sized theory of truth by synthesizing Tarskian semantics with a theory of reference. Each of them offers means for providing selective explanations of reference from within productivist guidelines, just as the theory of truth requires. But neither of them goes so far as to offer a general analysis of reference. That additional tall order is not their aim. For this reason, they both fall short of the prescription Field (1972) places on an account to complete his formula for a theory of truth.

If we want to see accounts of reference that make Field's conditions their own mandate, we must look to the next chapter in the history of theorizing about reference. The decades that followed saw a concerted effort to realize Field's vision. The naturalized content program of the eighties explicitly aimed to uncover a general account of reference of the kind that Field was looking for: namely, an IR account of reference cast in non-semantic terms. The key players in this effort were Stampe, Dretske, Fodor, Papineau, and Millikan.

Although the accounts from this period can be seen as descendants of their causalist predecessors, there was also a shift in focus. Whereas Kripke was concerned with the semantics of natural language expressions, the next generation of causal theorists—in keeping with the rise of cognitive science—were primarily concerned with the representational properties of intentional mental states. Thus, for them, the primary bearers of semantic features (i.e. reference, truth) were thoughts, beliefs, desires, and so on. Moreover, many of these authors postulated a language-like system of mental representations (a language of thought) to underwrite the intentionality of the attitudes. (Fodor 1975 and Field 1978 are the classic sources of this hypothesis.) According to this view, subsentential mental representations are the primary bearers of reference, and sentential mental representations are the primary bearers of truth. The semantic properties of public language, on the other hand, are (somehow) derivative of the semantic properties of thought.

Nonetheless, the Fieldian lure of naturalistic reduction was still in full swing. We find it in Fodor when he writes,

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of spin, charm, and charge will perhaps appear upon their list. But aboutness surely won't;

> intentionality simply doesn't go that deep. It's hard to see, in the face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves neither intentional nor semantic. If aboutness is real, it must be something else. (1987: 97)

As in Field (1972), the reason for demanding one's theory of reference (representation, aboutness) to be cast in non-semantic terms stemmed from a broader commitment to metaphysical naturalism and physicalism.

Besides looking for definitions of reference that could vindicate naturalism, the naturalized content theorists of the eighties also operated with another goal in mind. They also endeavoured to give *general* (i.e. necessary and jointly-sufficient) conditions for reference fixation. Should a theory deliver such conditions, it would then explain the selection of reference for specific terms by subsumption under general law. So the naturalized content theories do share the goal of offering selective explanations of reference. However, they approach these explanations from a place of general, overarching theory.[27]

Numerous attempts have been made to devise theories that meet these three aims: analyzing reference, naturalistic reduction, and selective explanation by way of general necessary and sufficient conditions. A list of the most well-known ones must include the following.

> *Causal theories of mental representation.* In his (1977), Stampe argues that the relation of mental representation is identifiable with a certain sort of causal explanation. Roughly, for x to represent y, according to this view, x's features must be causally explained by y's features.
>
> *Informational theories.* Dretske (1981) proposes that semantic relations (i.e. representation, reference) are understandable as a species of *informational* relations, where the latter is cashed out using the mathematical notion of information. Roughly, for x to bear information about y is for instances of x's type to reliably correlate with instances of y's type, such that we can extract probabilistic information about the instantiation of y's type by instantiation of x's type.

[27] Consider this passage from Stampe (1977): "The first of these [tasks for a theory of reference] may be thought of as the 'synchronic' dimension of the question, What determines that it is the particular thing of the relevant kind that is the object referred to or seen, and not some exactly similar thing of that kind? (What determines that it is twin A, and not twin B, that the photograph represents). That, I shall say, is determined by the fact that a system of causal relations of a certain (generic) kind—a kind constitutive of representation—connects one but not the other object to the relevant representation. This will define a causal relation from which the twin that took the picture is excluded." (44)

> Similarly, Fodor (1987, 1990, 2008) develops an 'asymmetric causal dependence' theory of mental representation. The basic idea here is that the mental representation x represents y just in case there is a law-like correlation between the tokening of x and the presence of instances of y. However, to block the theory from counting reliable misrepresentations as instances of representation, Fodor adds the further qualification that representing is a matter of *fundamental* correlation (1987: 101–10). So the reason why 'cow' doesn't represent *horses-in-dim-light* is that the correlation between 'cow' and *horses-in-dim-light* depends on the correlation between 'cow' and *cows*, and not the other way around.
>
> *Teleological theories.* Millikan (1984) and Papineau (1984) pioneered the idea that the representational features of the mind ought to be cashed out by appealing to evolutionary functions. The essential idea here is that a mental representation represents whatever it *evolved for the purpose of indicating.* The teleological notions in this claim then get chased out in terms of Darwinian selection.

Each one of these basic ideas has been developed into enormously sophisticated accounts. However, rather than exploring the details, I must instead offer some general comments about the program.

As the eighties came to a close, it became apparent how difficult it is to follow through with the project of naturalizing content if generality and reduction remain the aims. Any theory of reference that's intent on complying with the formula provided by IR must confront the fact that each representation bears a variety of causal relations to a variety of things, many of which are not the referent. They are thus obliged to sort out the content-conferring relations from the non-content-conferring ones, and the distinguishing factor must be specified non-semantically. (Fodor calls this the 'disjunctive problem' (1987: 102).) Each of the proponents has their own way of dealing with this problem, but without going into the details, it is fair to say that two further worries continuously plagued their efforts. For one, it appears that, in practice, the link between a particular representation and its referent will depend on particular factors that are specific to the representation in question.[28] This creates an obstacle to the *generality* of theory. Secondly, even if we can draw a general demarcation between content-conferring and non-

[28] There are a few different problems that I'm alluding to here. First, the link between a representation and its referent will typically be mediated by the subject's own theory of the subject matter. Since theories are also representational objects, their representational properties would also have to be accounted for if our aim is a general reduction of the reference relation. This creates a circularity worry; see Cummins (1997) for a full explanation of the problem. Secondly, the key proposals for naturalizing content variously appeal to *normalcy conditions* (Stampe), *development stages* (Dretske), *asymmetric dependence between word-world laws* (Fodor), and *evolutionarily conferred functions* (Millikan); and the crucial details of these things (which will matter to reference fixation) will vary depending on which representation is in question.

content-conferring relations, our application to specific cases can often be driven by semantic intuitions (Adams and Aizawa 2015; Loewer 1987). This creates a nuisance for the goal of reducing reference.

We will see in chapter two how these setbacks for the naturalized content program came to be a major liability for the object-based correspondence theory of truth. We have already seen enough to begin to appreciate the problem. I have argued that this conception of truth requires *some* answer to the questions posed by metasemantics. So if indeed those answers are tied to the success of an IR account of reference, then the failure to deliver one will threaten the whole conception of truth. Responding to this worry will be one of the main tasks of chapters two and three.

1.7 Truth-bearers

Until now, I have been fairly non-specific about truth-bearers. When discussing theories of semantic composition (§1.5), we focused on the sentences of artificial and natural languages. Then, in connection to the theories of naturalized content, we spoke of mental representations.

But a full-sized theory of truth should, ideally, address the pertinent questions as they pertain to each kind of truth-bearer: sentences, utterances, propositional attitudes, and perhaps the propositions themselves. So to complete this overview of the object-based correspondence theory, we must attend to these matters. However, since the topic is large, I can only offer the most cursory remarks.

The first point to make is that, once again, the general architecture of the theory places a rigid constraint on the permissible range of truth-bearers. Suppose that truth conditions are understood as a product of the theory of reference combined with principles of semantic composition. In that case, it follows that whatever can be true or false must be the kind of entity that is amenable to both.

In other words, the truth-bearers must be items that belong to a relatively *language-like* system of representations. It must be language-like in the sense that: (i) it contains syntactically simple parts that enter into symbol-object relations like reference, representation, and application (the purview of metasemantics), and (ii) the truth conditions of the sentence-sized representations are determined by their syntactic structure and the semantic features of their smallest parts (the purview of semantics). Given these constraints, it makes sense to focus on the linguistic case first since the semantics and metasemantics of language are not particularly murky when compared to mental or abstract truth-bearers.

Regarding the mental case, the object-based correspondence theory has several options. As long as it abides by these two constraints, it can be reasonably flexible or ecumenical about the nature of the truth-bearers (a point made by Glanzberg 2015).

For instance, because of these two constraints, the theory can naturally be seen as an ally to the representational theory of mind (Fodor 1975; Field 1978). According to this theory, a belief (thought, intention, desire, …) is a three-way relation between a subject, a mental representation, and a content. A mental representation is a syntactically structured, sentence-like vehicle of semantic content that is composed out of primitive symbols which are hypothesized to be within the architecture of the mind. These symbols and mental representations make up the so-called 'language of thought'; they are supposed to be the internal medium that facilitates mental states (e.g. belief) and mental processes (i.e. inference, via computation).

The representational theory of mind is an empirical hypothesis—a conjectured explanation of how the mind works. Nonetheless, if true, it would supply the object-based correspondence theory with a set of truth-bearers that are appropriate to intentional mental states. That is because mental representations are hypothesized to exhibit the right kind of structure for truth-bearers according to this conception of truth. They comprise a finite stock of syntactically simple symbols, many of which bear mind-world relations of reference, and the truth conditions of complex representations are determined by the semantic facts pertaining to the simples by semantic means. Let's say I believe *this is a piece of gold* as I hold a rock in my hand. According to the representational theory of mind, I then have a sentence processing in my mind (in a way that is characteristic of belief) which is an internal analogue of the sentence 'this is a piece of gold'. It is analogous precisely because it shares a similar subject-predicate syntactic structure and constituent symbols that mean *this* and *is a piece of gold.*

Although the object-based correspondence theory sits well with the language of thought hypothesis, its dependence on the latter may not be inevitable. But to properly explore this connection would take us too far afield. Suffice to say that if the merger between the two falls through, the object-based correspondence theory may have other options. Traditionally, the most common analysis of the intentional states (belief, desire, etc.) is that they are relations between a subject and a *proposition* (Hanks 2009). So we could instead account for truth for the attitudes by accounting for the truth of propositions. All we need is an account of propositions that exhibits the right structure. And indeed, there are options on the market that would be congenial to the object-based approach to conceptualizing truth. (See, for instance, the theories of structured propositions provided by King 2007 and Soames 2010.[29])

For my purposes, it is enough to mention that there are these various options for mental truth-bearers. However, I will not endeavour to develop this theory of truth with the specifics of

[29] King (2007)'s account of propositions is broadly Russellian in the sense that the proposition <Michael Swims> is a fact that is composed of Michael himself and the property of swimming itself (among other things). This means that the relation of sub-propositional *representation* between the <Michael> constituent and Michael would simply be identity. No further elaborate metasemantic theory is required, *for propositional representation.* But the overall theory of truth that incorporates this theory would still require an elaborate metasemantic theory as part of an explanation of how the sentence 'Michael swims' comes to express the proposition <Michael swims>.

these accounts in mind. My official policy is to remain neutral on the nature of mental truth-bearers, and I will continue to focus on the linguistic case. To justify this policy, I offer one reason for thinking that linguistic tokens are primary truth-bearers according to this conception of truth.

According to propositionalist orthodoxy, a sentence inherits its truth value from the proposition it expresses. Hence, the propositionalist will typically endorse:

> S is true(/false) in virtue of expressing P (in context) and P is true(/false)

(for sentence S and proposition P).

But according to the object-based correspondence theory, a sentence inherits its truth value from the semantic features of its parts, the rules of semantic composition, and the way the world is. Take, for example, the sentence S = ‘London is pretty’. S has the following attributes:

- ‘London’ refers to London
- ‘is pretty’ applies to pretty things (or expresses the property of prettiness)
- the subject-predicate form represents the attribution of the predicate to the referent of the subject term (or the attribution of the property *expressed by the predicate* to the referent of the subject term).

These attributes are all features *of the sentence and its parts*. We need not mention a distinct object, the *proposition,* to attribute these features to S. Moreover, according to the object-based correspondence theory, these attributes alone suffice to explain S’s truth condition: they entail that S is true if, and only if, London is pretty. Supposing that London is pretty, this yields that S is true.

Perhaps the best explanation for S having these semantic features is that S expresses the proposition that London is pretty. *But on the contrary, it seems to me that things are the other way around.* That is, S seems to express the proposition it does *in virtue of these semantic features.*

Take another sentence S*, that expresses the same proposition. Let S* be ‘Londres est jolie’. Why do S and S* express the same proposition? The answer (at least in part) must include the fact that they share the same semantic properties. Specifically, it is because ‘London’ and ‘Londres’ share the same referent, ‘is pretty’ and ‘est jolie’ apply to the same things, and their respective syntactical structures correspond to the same rule of composition. In short, the determination of propositional content *depends* on the semantic features of words.

This suggests that these semantic features (reference, predicate attribution, property expression, etc.) come *before* proposition expression in the order of explanation. That *S* *expresses P* is grounded in the semantic features of S, like reference (as opposed to the other way around: i.e. that reference depends on which proposition S expresses).

This raises the question: what are the primary bearers of *reference*? The answer, I take it, is that it is *tokens of words* produced by our utterances. We use words in particular acts of speech to refer to things in the particular occasions that we use them.[30] (This may also include the tokens of mental representations that are produced by our cognitive acts, if there are such things.) After all, every theory of reference (surveyed in §1.6.2) conceives of reference as a relation between words (including mental symbols) and things. Thus, our metasemantic theories treat words as fundamental bearers of the pertinent semantic features. We might gloss this by saying that *for metasemantics, linguistic reference is primary*.

Putting these thoughts together, we have: (i) linguistic tokens are a primary bearer of the pertinent semantic features, like reference, (ii) these semantic features suffice to explain truth, and (iii) these semantic features are prior to, and hence independent of, proposition expression in the order of explanation. Together, this justifies focussing on sentences rather than propositions as the bearers of truth and falsity. (None of this implies that there are no propositions. All it implies is that propositions need not be taken as fundamental from the perspective of an object-based correspondence theory.)

1.8 Conclusion

Returning now to our main theme: what *is* the nature of truth, according to this version of the correspondence theory? The answer that we have developed is that a truth-bearer is true if its truth conditions are met, and its truth conditions are explained as a product of the referential

[30] Following Strawson (1950), one might object that words, by themselves, do not refer to anything. It is a category mistake to attribute semantic features to words. Rather, it is *people* who refer to things by *using words* in the course of performing a speech act (Strawson 1950: 326). In response to this objection, I agree that semantically-imbued token expressions must be produced by a speaker's speech act (or, in the mental case, by a cognitive process). However, it seems to me that the objector is thinking of words as mere physical inscriptions on a page or sounds in the air. I am thinking of words as more than this. I am thinking of them as as *artifacts* of a speech act (or cognitive process) that are infused with semantic properties by virtue of how they are produced (e.g. the speaker's intentions, cognitive history, the linguistic division of labour in their speech community, etc.). Traditionally, speech acts have been decomposed into their *illocutionary force* (what type of act it is: an assertion, conjecture, question, etc.) their *content* (the thing specified by the that-clause; *what is said*), and the *phonetic act* (the act of producing such-and-such sounds) (see Austin 2018). I am suggesting that we should abstract from speech acts another kind of object that is produced: linguistic vehicles of content that are characterized (in part) by their semantic features.

Since I am thinking of words and sentences in this inflationary way, one might wonder what distinguishes them from propositions and their constituents. There are two things. For one, words are typically distinct from their referents, whereas this isn't the case for the constituents of propositions according to one highly compelling account of them (Russell's). For another, propositions *are* meanings; they do not *have meanings.* Words and sentences, on the other hand, *have* meanings, and they also have many other features besides. They also have syntactic features, a particular history, they are produced at a particular time and place, and by a particular speaker, with particular referential intentions, in a particular language, and so on.

properties of its simple parts (and the explanation of how these obtain), along with the rules of composition for its syntactic type. Now, despite the superficially truistic appearance of this answer, it places several non-trivial constraints on our overall picture. To reiterate:

I. The truth-bearers must belong to a system of representation that admits of simple and complex syntactic structures, referential properties for the simples, and truth conditions by composition for the complex structures.

The least that we can say is that Tarskian artificial languages fit the requirement precisely because they were designed to. In addition, large fragments of natural language will also (plausibly) fit this requirement, but explaining how this is so is the purview of semantic theory and natural language metasemantics. There may also be other systems of symbolic representation (i.e. mental representations, propositions) that are apt for truth and falsity in this picture. Still, in the interest of keeping this project self-contained, I must adopt a policy of neutrality on the nature of mental truth bearers.

II. The rules of semantic composition for natural languages that follow the structures given by generative syntax is an ongoing empirical investigation. There are several outstanding questions as to the exact nature of these rules, but we can see the efforts to deal with them as an ongoing and successful project.

III. This theory of truth depends on a theory of reference. Moreover, *reference* must come before *truth* in the order of explanation, according to the overall shape of this theory. Consequently, this means that reference must be explained atomistically (as opposed to holistically), along productivist lines.

This picture of truth may seem to be writing a long list of IOUs. It claims that the *entire* understanding of truth must rest on the further results to be obtained by semantics and metasemantics. It also makes an additional promissory note about truth for the intentional attitudes—especially if it is wedded to the representational theory of mind. Given all of this promise-making, the main theoretical choice is whether we can tolerate this large amount of debt for a theory of truth. Naturally, the opponents of this theory will find it intolerable. In the next chapter, we'll consider the attempt to get away with less (also known as the *deflationary* theory of truth).

On the other hand, it is sometimes said that philosophy is the handmaiden to the sciences. Ideas that are initially philosophical can sometimes (when successful) give way to full-fledged sciences. Now, when it comes to this version of the correspondence theory, its adherents would want to tell a story that fits this theme. In their view, theorizing about truth begins with vague intuitions about truth being a matter of correspondence with reality. But since these intuitions are vague, improvements are required. Then, after dabbling in the attempts to cash out these ideas in the metaphysics of propositions and facts, it became apparent that the best way to proceed is through semantic and metasemantic theory (broadly construed). But the relevant areas of inquiry are no small tasks. Our understanding of these areas is still developing. Hence the current state of the object-based correspondence theory.

Chapter 2: Between deflationism and inflationism: a moderate view on truth and reference[31]

2.1 Introduction

Deflationism of all kinds is fashionable these days, and no kind is more fashionable than the deflationary theories of *truth.* By an ironic twist of fate, this trend was started in the early nineties by Hartry Field himself, along with Stephen Leeds and Paul Horwich. After spending a couple of decades advocating for the object-based correspondence theory, Field became dissatisfied with the agenda that he had set and converted to the most formidable rival of his former view. (We will discuss his motives in this chapter.) Deflationism is now described by some authors as the 'near orthodox' position on truth.[32]

In its formative years, the deflationary view was generally fostered by skepticism towards the prospects of an *inflationary* account of truth, particularly along the lines of the view presented in chapter one. The early authors were specifically doubtful that truth could be explained on the basis of an inflationary theory of reference. Instead, they argued that both truth *and reference* ought to be elucidated by mere trivialities—such as *'snow is white' is true iff snow is white* and *'Kilimanjaro' refers to Kilimanjaro.*

What deflationism has to say about truth has now been fairly well-explored in the literature, but its claims about reference have received relatively less attention. There is a certain irony about this, considering that the issues surrounding reference were central to deflationism's initial motivations. It is thus my aim for this chapter to focus on reference and bring these issues out to the fore.

One of my main contentions is that the debate between deflationism and inflationism can be profitably recast as a debate over what it takes to explain a reference relationship. Each of these views represents a distinctive approach to explaining why a given word has its particular referent. Moreover, once we see the debate in this light, we find that there are multiple points of contrast between the two approaches. For one, they each have different *orders of explanation*—different prescriptions for *what is explaining what* when it comes to semantic phenomena. For

[31] This chapter is published as Moore, G. S. (2022) 'Between Deflationism and Inflationism: A Moderate View on Truth and Reference', *The Philosophical Quarterly*, 72/3: 673–94. It is reprinted here with permission. The present version differs from the published version only in that it includes an additional example and some additional framing to smooth out the traditions between chapters. Also, a few paragraphs from the published version were moved to the previous chapter.

[32] E.g. Simon Blackburn (2006: 249); Amie Thomasson (2014a: 185).

another, the inflationist is typically taken to be beholden to a reductive explanation of reference, whereas the deflationist is doubtful of this project.

My second main contention is that these two points of contrast need not come together to exhaust the space of possible views. There is room for a plausible middle ground: a moderate version of inflationism. My moderate inflationism will reject the deflationist's structure for explaining reference, so it counts as genuinely inflationary. However, it also rejects the reductionist ambitions of the earlier inflationists, so it isn't an apt target for deflationary skepticism.

Within the context of my broader project, the results of this chapter shed light on a problem that was left unresolved in chapter one. There it was argued that an object-based correspondence theory requires *some* metasemantic explanation of the primitive semantic facts, but besides the general productivist constraint, it was left open-ended as to what other forms this explanation must take. In this chapter, we examine an extreme minimalist approach to accounting for the semantic facts. By seeing what it's missing, we thereby uncover the minimal requirements for the metasemantic component of an object-based correspondence theory.

2.2 Deflationary accounts of truth

Although the deflationary theory of truth is not our primary focus, it can serve as a natural starting place for the deflationary theory of reference.

When it comes to truth, a typical way to summarize deflationism is to say that we don't need a 'deep' theory to understand what truth is. That is because, according to deflationism, all it takes for a sentence '*p*' to be true is simply for it to be the case that *p*. Apart from this, there isn't anything more to say to explain '*p*'s truth. In particular, we don't need to say that truth consists in correspondence or provide any additional explanation of how '*p*' relates to the world.

All of this can be made more precise by spelling out two core deflationary theses.[33] The first one is a positive claim about the role of the truth predicate and the second one is a negative claim about the metaphysics of truth.

According to deflationism's first thesis, the primary reason for having a truth predicate in our language is to fulfill certain logical or syntactic needs. (Its purpose is not to refer to a substantive property.) To be specific, the role of the truth predicate is to provide a means for swapping a sentence 'S' with an equivalent sentence '"S" is true' (nominalization), and for taking quotation marks off a quoted sentence (disquotation). This turns out to be incredibly useful for

[33] Given the diversity of deflationist positions, there is always a hazard in claiming that any particular thesis is amongst the 'core'. A more qualified statement is that these theses are central to the authors that I am concerned with. The disquotationalist view of Leeds and Field straightforwardly endorses both of them. Horwich's minimalism also implies them, but in a roundabout way which I will explain shortly.

various expressive purposes. It allows us to quantify over a plurality of sentences and state that each one of them is true, which ultimately has the effect of stating each one. Deflationists commonly gloss this feature of the truth predicate by saying that it is a device for expressing generalizations. In their view, this expressive function is the chief purpose of the truth predicate.

If the deflationists are right about the truth predicate's purpose, then they have a justification for why truth doesn't need a deep account. Basically, in order for the truth predicate to perform its essential function, it must be that an ascription of truth to a sentence is logically equivalent to that very sentence. This means that each instance of the following *disquotational truth schema* DT must hold:

(DT) 'S' is true if and only if S.[34]

The instances of this schema are called *T-sentences* and they are foundational to the theories of Field (1994a) and Leeds (1995). The deflationary explanation as to why these T-sentences hold is that they follow from the logic of the truth predicate *as a device of disquotation.* For this reason, deflationists often claim that the T-sentences are 'trivial', or 'conceptual', or 'analytic'.

This brings us to the deflationist's second thesis. Since they claim that the logic of 'true' explains the T-sentences, the deflationist must deny that these sentences admit any *deeper, more substantial* explanation. Consider the T-sentence, *'snow is white' is true if and only if snow is white.* For the deflationist, this equivalence holds in virtue of the logic of the truth predicate, which is explained by its expressive function. In that case, there shouldn't be any room for any further explanation as to why the sentence 'snow is white' has this truth condition. There shouldn't be any further explanation in terms of the relations that 'snow is white' bears to its subject matter, or in terms of anything else that could constitute a metaphysical theory of truth. For the deflationist, truth is not the sort of property that has some hidden underlying nature that awaits our discovery.

If truth can't be given a deep characterization, then what *can* the deflationist say about the property of truth? The most that they can offer is the schema DT. In their view, all of the philosophically important facts about the property of truth are given by the instances of DT. They can thus take the T-sentences to collectively define truth. Field (1994) calls this the *pure disquotationalist theory.* This pure disquotationalist theory will represent, for the deflationist, the core set of facts about truth.

Although DT gives us the core of the theory, there are several well-known complications that prevent the deflationist from simply ending their story here. The disquotational schema, by itself, really only works in certain circumstances. The instances of DT are only guaranteed to

[34] This must be restricted to the instances where the 'S' mentioned on the left-hand side has the same meaning as the 'S' used on the right-hand side. There will be more on this qualification shortly.

hold when the ‘S’ mentioned on the left-hand side can intuitively be said to *mean the same thing* as the ‘S’ used on the right-hand side. But if that congruence breaks, then all bets are off. Take, for example, a context-sensitive sentence that is taken out of context: e.g. ‘I am a beekeeper’, as uttered by some person who isn’t me. To ascribe the right truth conditions to their utterance, I mustn’t invoke the T-sentence, *‘I am a beekeeper’ is true if and only if I am a beekeeper*. For another source of problems, consider the hypothetical scenario where a given sentence ‘S’ is used to mean something different from what it actually means. What *would* ‘S’s truth conditions be in that case? Again, DT is not going to provide the right answer.

The first step to handling these problems is to restrict the pure disquotationalist theory to a special set of sentences. To be specific, a person can invoke a pure disquotational theory only for the sentences *that they understand* (that are expressible in their language) and *are suitable for use in their context* (Field 1994a: 250, 279–81). We can call these sentences—the ones that are suitable for disquotation—the ones that belong to a *home language.* Since a home language is designed to take care of all kinds of context-sensitivity, each person’s home language will be highly specific to them and their context. Field (1994a) suggests that it can be identified with their internal system of mental representations—their *language of thought*.

Following this, deflationism can then deliver the truth conditions for the sentences outside of one’s home language by expanding its resources. Let ‘S’ be such a sentence (perhaps ‘S’ is context-sensitive and taken out of context, or it is considered with a counterfactual meaning, or perhaps ‘S’ is simply foreign to us). We can understand an attribution of truth to ‘S’ by first *translating* it or *interpreting* it using our home language, and then invoking the pure disquotational theory. For example, I can interpret another person’s use of ‘I am a beekeeper’ as ‘*he* is a beekeeper’, and then invoke the relevant T-sentence. The result is that ‘I am a beekeeper’ (as spoken by *him*) is true if and only if *he* is a beekeeper. In general, the deflationist invokes a two-step approach to capturing the entire range of truth attributions:

> (EDT) ‘S’ is true if and only if there is a sentence ‘P’ such that ‘P’ interprets/translates ‘S’ and ‘P’ is (disquotationally) true (where ‘P’ is in my home language and adjusted to my context).[35]

This *extended* theory of disquotational truth still upholds the deflationary idea that the explanation of truth ought to bottom out in the instances of disquotation given by DT. It’s just that, in some cases, in order to utilize the pure disquotationalist account of truth, we must first interpret or translate the target sentence into something more suitable.

This two-step approach will naturally invite questions about how this translation/interpretation step is supposed to work. *What are the rules for interpretation? What facts explain*

[35] ‘EDT’ is for *extended disquotational truth,* to use the name given by Field (1994a).

why a given translation is appropriate? We will soon come to see that these issues form the crux of the deflationist's view on language-world relations. But before we attend to their answers, there are a few more details of the deflationist's picture to consider.

Some deflationists—notably, Paul Horwich—prefer to take *propositions* as the primary bearers of truth. According to them, the most fundamental schema in our theory ought to be:

(PD) <*P*> is true if and only if *P*.[36]

For them, it is this schema, rather than DT, that supposedly explains the expressive role of 'true', as per the first deflationist thesis. It is also this schema that is supposedly fundamental to characterizing the property of truth.

Despite the noticeable difference between this theory and the previous one, it is important to observe that, even for Horwich, his account of sentential truth ultimately ends up being deflationary in my intended sense.[37] This is because he also endorses the second core deflationary thesis when it comes to sentential truth. To see this, first note that sentential truth can be defined in terms of propositional truth:

For all S, S is true iff there is a proposition x such that S expresses x and x is true.

In addition to this, Horwich gives a 'deflationary' account of what it is for a sentence to express a proposition. Basically, for him, the expression relation is not to be cashed out in terms of a robust account of the content-determining relations between a sentence and its subject matter. Instead, it is characterized by another trivializing schema:

'S' expresses <P> if and only if 'P' translates/interprets 'S'.

(Once again, the translation/interpretation of 'S' into 'P' functions to translate 'S' into our home language and adjust for context-sensitivity if needed.) When we combine these two principles together with PD, we essentially recreate EDT (Horwich 1998a: 101–2). And in the special case where 'S' is in our home language, these principles entail the T-sentence, *'S' is true iff S.* The upshot is that, even for Horwich, the explanation of a sentence's truth conditions ends up being a matter of interlinguistic translation and the logical workings of the truth predicate.[38] He too must

[36] 'PD' stands for *propositionalist deflationism*. The '<…>' notation means '*the proposition that …*'.

[37] Some theorists (e.g. Scott Soames 1999 and Matthew McGrath 1997) take a deflationary stance on propositional truth and an inflationary stance on the relations between sentences and propositions. As a result, their views do not have the deflationary implications for sentential truth and reference that will be my main focus here.

[38] Along with an extra syntactical step for converting quoted sentences into names for propositions.

deny that sentential truth is explicable by a more substantial relationship between language and the world.

We can summarize the common deflationary thread by considering what *explains* or *grounds* the fact that a certain sentence ‘S’ has the truth conditions that it has. The deflationist claims that the explanation for ‘S’s truth conditions must ultimately bottom out in the trivial logical features of the truth predicate. In the special case where ‘S’ is primed for disquotation, we may recite *‘S’ is true iff S*; this is alleged to be a fact about the logic of ‘true’. If, on the other hand, ‘S’ isn’t in our home language, then it needs to be translated/interpreted into something that is. This step is *interlinguistic* since it maps linguistic items to linguistic items. Once that’s achieved, we can then appeal to the same trivial schema as before. So all in all, truth is *not* ultimately explicable in terms of language-world relations.

2.3 Contrast 1: the order of explanation

This sets the stage for the deflationist theory of reference. In order to bring our discussion down to earth, let’s centre it around a couple of mundane examples of reference relationships. Doing so will allow us to tease out the first subtle contrast between deflationary and inflationary views.

Case 1 (Demonstrative) One very foggy night, my partner and I take a long walk on the beach. In the distance, there’s a dim light, mostly obscured by the fog. My partner looks at it and points towards it and says ‘*that* is a lighthouse’.

Case 2 (Proper name) Continuing on our walk, our conversation turns to natural wonders. My partner says in a matter-of-fact way, ‘Kilimanjaro is the world’s tallest free-standing mountain’.

In each case, it is clear what the referent should be. In Case 1, my partner uses the term ‘*that*’ to refer to a certain object—let’s call it *o*—which is the cause of the dim light. Her utterance is true provided that *o* is, in fact, a lighthouse. Case 2 is even more obvious: ‘Kilimanjaro’ refers to *Kilimanjaro* (the famous dormant volcano in Tanzania). Her utterance is true provided that Kilimanjaro is the tallest free-standing mountain; which it is, so her utterance is true.

Now suppose that our task is to *explain* these facts of reference. That is, we want to provide a metaphysical explanation as to why my partner’s use of ‘*that*’ refers to *o*, the cause of the light, rather than (say) the distant mountains; and we want to explain why ‘Kilimanjaro’ refers to Kilimanjaro, rather than any other mountain, natural object, or artifact. What should we say?

Here is a sketch of a fairly natural answer. But as we will see shortly, it is only available to the inflationist; the deflationist cannot accept it at face value. (To be clear, this isn't the only kind of explanation that falls under the inflationist heading. Endorsing this explanation is sufficient, but not necessary, for departure from deflationism.[39])

Inflationary explanation 1 (Demonstrative). First, when my partner uttered '*that's* a lighthouse', she did so with a certain intention in mind. She had an intention to refer to the object that she perceived in the distance. Since it is *o* that she's perceiving, *o* is the referent of her expression. If we really wanted to get into the detail, we could explain why her intention and perceptual state are about *o* by appealing to a certain causal link between her intentional states and the object. We could then outline a causal chain, C_1, that relates her utterance to her cognition, and her cognition to *o*. In that case, her token '*that*' refers to *o* because it is connected by C_1 to *o.*

Inflationary explanation 2 (Proper name). Secondly, when my partner said 'Kilimanjaro is the world's tallest free-standing mountain', she did so with the intention of speaking about Kilimanjaro. Now, although she has never seen the mountain herself, she has some store of information concerning the so-called 'Kilimanjaro' that she has gathered through various sources (books, encyclopedias, hearsay). Some of these sources may be first-hand witnesses, or maybe they all aren't. But either way, if we trace back the passing of information from source to source we will eventually find some original 'Kilimanjaro'-users that are first-hand witnesses of the mountain. Let's call the socio-historical chain of information-passing that traces back to the mountain C_2. The explanation, then, is that her use of 'Kilimanjaro' refers to Kilimanjaro because it is related by C_2 to Kilimanjaro.

This kind of explanation will be familiar and congenial to those who have been brought up in the Kripke-Putnam tradition. For our purposes, it doesn't really matter what the details of the causal links are. All that matters is that there are some causal pathways like C_1 and C_2, and they can be vaguely gestured towards as part of an explanation of these reference facts.

As I have said, the deflationist cannot accept either of these explanations as they stand. It is important to see why this is. The central reason stems from their negative metaphysical commitments towards truth. To put it briefly, *if* the facts of reference were explicable along these

[39] The two other notable kinds of inflationism are *interpretationism* and *primitivism.* As noted in chapter one, the interpretationist (e.g. Donald Davidson) takes reference to be determined by the best overall interpretation of the subject's verbal behaviour. A primitivist about reference would take the reference facts to be brute and fundamental. I mention these alternatives only to highlight the fact that the kind of inflationism discussed in this chapter does not exhaust the available alternatives to deflationism. But it is, nonetheless, the alternative that I favour, and it is the kind of inflationism that the deflationists were reacting to, as we'll discuss shortly.

lines, then that would threaten the deflationary claims about truth.[40]

To see this, consider the fact that truth is interdefinable with reference and predicate satisfaction (at least for the first-order fragment of our language, as shown by Tarski; see §1.5.1). It follows that if there are any inflationary explanations of reference, then they can serve as the basis for an inflationary explanation of truth. One could explain the truth conditions of sentences by combining a theory of semantic composition (say, the Tarskian recursive clauses) with the inflationary explanations of reference, following the template of chapter one.

Suppose, for example, that we wished to explain why my partner's sentence '*that's* a lighthouse' has its truth condition. The inflationist can proceed by first explaining why '*that*' refers to *o* and why '*is a lighthouse*' refers to lighthouses, along inflationary lines (e.g. Inflationary explanation 1).[41] They can then explain how the sentence's truth condition is determined by articulating the rule of semantic composition for subject-predicate sentences. The result will be an informative account of why '*that's* a lighthouse' is true if and only if *o is a lighthouse.*

But as we have seen, this is exactly the sort of account that the deflationists reject. They claim that a sentence's truth conditions cannot be explicable by anything more than translation and disquotation. So for that reason, they must also reject the inflationary explanations of reference and instead opt for something more deflationary.

The deflationist theory of reference is closely parallel to the deflationist theory of truth. Much like truth, the deflationist claims that reference is, in a certain sense, 'insubstantial'. By this I mean that their full account of reference will reside in the alleged logical features of 'refers', which are displayed by the *reference deflationist* schema:

(RD) '*a*' refers to *a* (if *a* exists).

[40] This claim is fairly uncontroversial in the literature. It is accepted by each of the deflationists that I'm citing: Field, Horwich, and Leeds. Horwich (1998a) and Thomasson (2014a) give a sketch of my reasoning here; see Taylor (2017, 2020) for an in-depth discussion of the connection between truth deflationism and reference deflationism.

[41] I'm assuming that the application of predicates can be counted as a reference relation and that the inflationary explanations for singular reference have analogues for predicates. However, strictly speaking, we don't need to make this assumption in order to make the current point. The deflationist's claims about truth would be just as threatened if the only kind of inflationary explanations of reference pertained to singular terms. To see this, take any subject-predicate sentence, '*a is F*'. Tarski's theory will deliver the equivalence: '*a is F*' is true iff the referent of 'a' is F. (Assume that the semantic contribution of 'F' is given either by inflationary means or deflationary means.) Now, suppose that it is possible to give an inflationary explanation as to why the referent of '*a*' is *a* (say, in terms of causal relations borne between the tokens of '*a*' and *a*). In that case, we would have an inflationary explanation as to why '*a is F*' has the truth condition that *a is F* (as opposed to the truth condition that *b is F* for any $b \neq a$). Such an explanation is exactly what is denied by deflationism.

(As before, this schema is restricted to the singular terms of our home language.[42]) For the deflationist, the instances of this schema are supposed to be the last stopping point for explaining the facts of reference. This further implies that reference relationships are not to be explained by any more fundamental relations between language and the world. The fact that *'Kilimanjaro' refers to Kilimanjaro* (as used in my own home language) is not to be explained by any further facts about my usage of the word 'Kilimanjaro' and its relation to Kilimanjaro. The view explicitly denies the possibility of a causal or descriptivist theory of reference (Field 1994a: 261–3; Leeds 1995: 15; Horwich 2005: 184).[43]

Much like the situation with truth, this cannot be the deflationist's entire story about reference, for the familiar reasons concerning context-sensitivity, counterfactual meanings, and the attribution of reference to foreign expressions. Again, they must expand the account to deal with these cases. And once again, their trick is to first translate/interpret the expression into something that's suitable for use in the home language and then apply RD. As a result, the full deflationary account of reference is given by another two-step schema:

> (*Extended Reference Deflationism/*ERD) If '*b*' translates/interprets '*a*', then '*a*' refers to *b* (if *b* exists).

As before, the term '*b*' is an expression from our home language which can substitute for foreign expressions or adjust for context-sensitivity if needed.

Returning to our examples, we now have some idea of how the deflationary explanation would go in each case. Ultimately the explanation must abide by ERD. Supposing that I'm the one giving this explanation, I must factor it into two steps: I must interpret/translate the terms into my home language and then disquote.

Deflationary Explanation 1 (Demonstrative). Given the circumstances and the way in which my partner is using her token of '*that*', her token '*that*' translates as my terms '*o*' and '*the cause of the light*'. Moreover, '*o*' *refers to o* and '*the cause of the light*' *refers to the cause of the light.* Hence, the token '*that*' refers to *o, i.e. the cause of the light.*

Deflationary Explanation 2 (Proper name). Given the circumstances and the way in which my partner is using 'Kilimanjaro', her term 'Kilimanjaro' translates (homophonically) to my term 'Kilimanjaro'; my term 'Kilimanjaro' refers to Kilimanjaro; hence her term 'Kilimanjaro' refers to Kilimanjaro.

[42] For present purposes, it is best to interpret RD and ERD (defined below) as applying to *tokens* of singular terms.

[43] See Taylor (2017: 48–58) for an elaboration of reference deflationism.

So far this gives us a general feel for the deflationary approach to explaining reference. Disquotation is the kingpin and translation is the accomplice. Evidently much relies on this notion of translation/interpretation, so to give the full story, we need to say more about how this works. For the sake of brevity, let's just call it *translation* to cover both cases.

There are a few things to notice about the peculiar way in which the deflationist must understand translation. For one, it is supposed to deliver a mapping between linguistic items (words and sentences) to other linguistic items (words and sentences in my home language) for the purpose of disquoting the result. It is thus an interlinguistic relation. It is not the same thing as an assignment of objects to expressions. It is a relation between words and words—not a relation between words and things.

Secondly, it is important to emphasize that translation, in the sense that's relevant to the deflationist, is always a mapping of words *into one's home language.* Remember, an individual's home language is a special-purpose language, specific to their context, that is designed to provide context-adjusted expressions that are understood by the subject and suitable for use. These languages are thus best thought of as individualized, context-specific idiolects. We shouldn't think of them along the lines of common cultural languages like English, German, or Mandarin. For this reason, we shouldn't think of translation as quite the same thing as translating, say, German into English. Instead, think of it as translating another person's speech into your own interpretation of it.

Lastly, it is important to observe that the deflationist appeals to translation *prior* to their (disquotational) explanation of reference and truth conditions. For them, it is partly *because* '*that*' translates as '*o*' that '*that*' *refers to o.* This places a steep constraint on how the deflationist can understand the workings of translation. Specifically, the deflationist cannot say that two expressions are translatable on the basis of shared reference or truth-conditional content. They cannot say that '*that*' translates as '*o*' because the two refer to the same thing, on pain of circularity. So whatever explains translation, it cannot include reference or truth-conditional content.

The deflationist thus owes us some explanation of how translation works. They need to tell us what makes a pair of terms intertranslatable if it is not for shared reference. For our examples, they need to explain why my partner's term '*that*' translates as '*o*' and her 'Kilimanjaro' translates as my 'Kilimanjaro', and their explanation cannot presuppose that the pairs are co-referential.

As it happens, both Field and Horwich offer solutions to this problem that are broadly similar in outline. Both of them claim that the meaning of an expression can be understood in terms of certain features of its use. A pair of terms would then be intertranslatable if they share the same meaning—that is, they share the relevant features of use. So, roughly speaking, when I translate a term '*a*' into my term '*b*', I am judging that '*a*' is used by its speaker in their context in much the same way (and in much the same circumstances) as I would use ‘*b*’ in my context. Again, ‘sameness of use’ must be cashed out without appeal to reference.

To see how this works in practice, we need more details as to which features of use are relevant to meaning and translation. And although our deflationist authors have divergent takes on this, what matters for our purposes is what they have in common.[44] For instance, both Field and Horwich list the *conceptual role* of a term as a potential constituent of its meaning (Field 1994a: 253; Horwich 1998a: 93–4). They also both allow for certain mind-world relations to enter the picture. Specifically, they highlight that a subject’s tokening of an expression can bear a law-like correlation with the presence of an object (or kind of object) in their environment. Field calls these ‘indication relations’. For example, I tend to think ‘cow’ whenever there are cows nearby; for this reason, my expression ‘cow’ will indicate cows. Both Field and Horwich claim that these correlations can serve as constituents of the meaning of an expression; and so two terms can mean the same thing in virtue of indicating the same things (Field 1994a: 254; Horwich 1998a: 93, 1998b: 45–6). Field (1994a) even observes that these indication relations are often explicable by causal relations between words and the objects they indicate (261–3). So even causal relations can enter into the deflationary picture of meaning.

In addition, Field (1994) also includes social facts into the potential ingredients of meaning (255–6). When we translate someone’s use of an expression, we are thus free to take note of how that expression has been used throughout the subject’s linguistic community. In particular, we may note the expression’s conceptual role for other speakers, its indication relations for other speakers, and its history of transmission in the community.

Let’s now return to our main examples. The remaining question concerns the translations of my partner’s terms into mine: which features of use can justify the translations? For present purposes, the important thing to notice is that deflationism allows for certain word-world

[44] The main points of disagreement concern the systematicity of theory and the notion of synonymy. Among the deflationists that we’re considering, Horwich is the keenest to develop a deflationist-friendly theory of meaning. His basic proposal is that the overall use of each meaningful expression is governed by a certain basic ‘regularity of use’. Since, for Horwich, the basic regularities are constitutive of meaning, two terms will have the same meaning if, and only if, they share the same basic regularity of use. This means that, for Horwich, there is always a right answer as to whether two terms are synonymous and hence intertranslatable. He writes that “there does exist a fact of the matter. Either two words *are* properly intertranslatable, or they are not—even though it may be impossible to say which is so” (1998a: 96).

Field, on the other hand, is not so optimistic. Without the help of a prior notion of reference or truth conditions, he hesitates to commit to a notion of meaning that can induce an absolute synonymy relation between the idiolects of different subjects (1994a: 271–4). In his view, the deflationist should be open to the possibility that there is no absolute interpersonal synonymy. In that case, there would be no unrelativized notion of a ‘correct’ translation between languages. Instead, there would be better or worse translations, and the standards for translation would be relative to the goals of the translator (the person ascribing truth conditions).

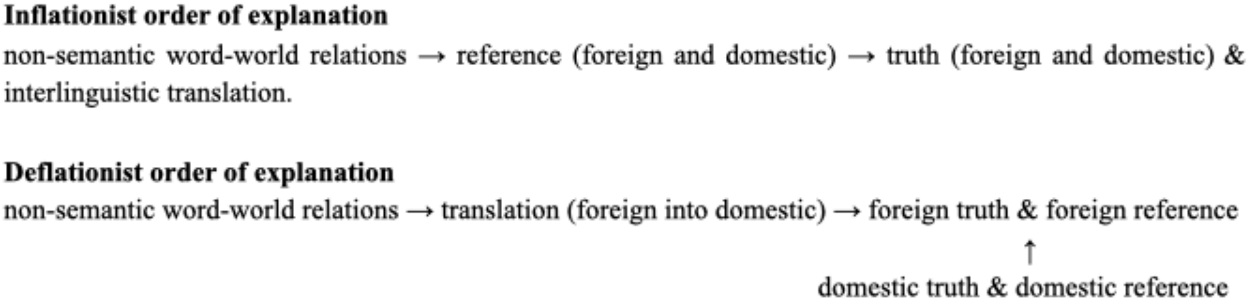
On this issue, Leeds agrees with Field. He writes that the deflationist “need not give a general account of what the standard translations have in common: perhaps there are some features of the use of each language that they preserve better than any alternative, but perhaps not… the *existence* of the standard translations are not in doubt, whatever one might think about the prospects for giving a general theory about them” (1995: 7). So in Leeds’ view, the deflationist is free to appeal to the fact that we *can* translate between languages. However this happens, we are (allegedly) able to translate without the help of an inflationary theory of reference or truth.

relations to characterize an expression's use, and thus figure into the explanation of meaning and translation. This includes correlations between utterances and objects, and it may even include the term's causal precedents and historical uses within a community.

For these cases, it seems like the most sensible thing to do is point to the circumstances in which each term was produced. Specifically, '*that*' translates as '*o*' because '*that*' bears a certain causal relation—C_1—to *o,* and my term '*o*' bears a similar causal relation to *o*. Her 'Kilimanjaro' translates as my 'Kilimanjaro' because her 'Kilimanjaro' has a certain casual history—C_2—and my term has a relevantly similar causal history. At any rate, the deflationist is certainly *permitted* to justify their translations in this way. To borrow a phrase from Field, these causal histories are "there to be observed; and a deflationist is as free to take note of [them] as anyone else" (1994a: 254). The only restriction is that they mustn't take these causal relations as *directly* explaining reference.

As we can see, the deflationist can fill their metasemantic stories with details that will make them sound very much like inflationists. Each party acknowledges the same non-semantic relations between the speaker and the world, and they may appeal to the same relations at some point or another in their overall accounts.[45] In the extreme, it is even possible for the deflationist to appropriate and mimic any explanation of reference given by the inflationist. Whenever an inflationist says "'*a*' refers to *b* because …", the deflationist can copy the "..." part and reply "that's why '*a*' is *translated* as '*b*'; reference remains disquotational". Wherever the inflationist sees grounds for reference, the deflationist can see grounds for translation.

Even though this is possible, the two views would still remain distinct. Despite the fact that they may use the same non-semantic relations as ingredients in their stories, they would still differ in their orders of explanation. Here is a depiction of *what's explaining what* according to the deflationist and the kind of inflationist that was introduced earlier:[46]



Inflationist order of explanation

non-semantic word-world relations → reference (foreign and domestic) → truth (foreign and domestic) & interlinguistic translation.

Deflationist order of explanation

non-semantic word-world relations → translation (foreign into domestic) → foreign truth & foreign reference

↑

domestic truth & domestic reference

[45] Maddy (2007) makes the same observation; "The disquotationalist I'm describing here is no less concerned about word-world connections than the correspondence theorist. In fact… the two may well be focused on precisely the same word-world connections" (163–4).

[46] As per footnote 39, this 'inflationist order of explanation' doesn't encompass every kind of non-deflationary picture in the literature. It is only representative of the productivist orientation (§1.6.1). Such a picture is sufficient, but not necessary, for a view to stand in opposition to deflationism.

Of course, when it comes to domestic truth and reference, the deflationist appeals to the alleged logical fact that ‘true’ and ‘refers’ are disquotation devices.

So here we have a key diagnostic to test whether a view is inflationist or deflationist. *Do (non-semantic) word-world relations explain reference relations directly, regardless of which language the terms belong to? Does the correct translation between terms depend on them referring to the same thing?* If a theorist answers ‘yes’ to either question, then that qualifies them as an inflationist. If, however, they’d rather explain interlinguistic translation in terms of non-semantic features of use, without mentioning reference, then their view harmonizes with deflationism.

2.4 Contrast 2: the nature of reference

Now that we have articulated these two different structures of explanation, I think it is fair to say that the inflationary one carries more prima facie appeal. Here are a couple of considerations that point in its favour, given what we have seen so far. (I am not claiming, at this stage, that these arguments are decisive.)

First, the deflationist's explanatory route seems roundabout and backwards. Take the demonstrative case for consideration. What is more plausible? Is it that my partner’s use of ‘*that*’ refers to the object *o* because of her referential intentions and her perceptual relations to *o* (and my translation of her term is correct *because* I get her intentions right)? Or does ‘*that*’ refer to *o because* I translate it as ‘*o*’ on the basis of my similar perceptual relationships? It seems that the deflationist gets things the wrong way around. Especially for demonstratives, it seems that reference has all to do with the production of the term and nothing to do with *my* translation.

Not only does the deflationary order of explanation appear to be circuitous, but it also carries some bizarre consequences. *Question: if a demonstrative is uttered in a forest and we’re not there to translate it, does it still make a reference?* For the inflationist, the answer is clearly *yes*, provided that the speaker bore the appropriate perceptual and intentional relationships to an object. Foreign reference, in their view, doesn’t depend on *our* translation. But for the deflationist, things aren’t so simple. According to ERD, reference is attributable only if the demonstrative is translatable into my home language.[47] But what if there isn’t any translation

[47] See Moore (2020) for an extended discussion on this point. Here is what Field (1994a) says about the deflationary understanding of reference for indexicals: “When I say that I ‘associate values’ with an indexical, of course, what I do is associate a mental occurrence of one of my own expressions (possibly itself indexical) with it. If I can’t associate a term with an indexical in a sentence, then I can’t attach disquotational truth conditions to the sentence.” (280)

(that is, my home language contains no term for the intended referent)?[48] Would it follow that I cannot say that the lone speaker has referred (from my perspective, confined to my own home language)? Maybe the deflationist can make sense of the intuitive idea that I can attribute reference in this case; but then again, the fact that this is puzzling for them shows that they're in an awkward position.

So, given that there is some pull towards the inflationary structure of explanation, and given that the two may end up focussing on the same non-semantic word-world relations at some point or another, one has to wonder: what is the appeal of deflationism? What does deflationism have to offer that inflationism doesn't? This will bring us to the second difference between the two views.

When the proponents of deflationism argue for their theory against inflationism, they invariably declare their skepticism towards the prospects of an inflationary theory of reference. Here is what our deflationists have to say.

> Part of the deflationary position, as I see it, is that the reference relation is very unlikely to have any… underlying nature. (Horwich 2005: 192)

> It is hard to see how the conditions for a deeper account of [the instances of RD] could possibly be satisfied. For a decent explanation would have to involve some unification, some gain in simplicity… Therefore we can conclude that the reference relations are not constituted by some more fundamental non-semantic relation. (Horwich 2005: 191)

> The project of giving anything close to a believable reduction of truth conditions [e.g. via an account of reference] to talk of indication relations is at best a gleam in the eye of some theorists. (Field 1994a: 255)

> Consider "rabbit": an inflationist presumably thinks that the set or property that my term "rabbit" stands for is determined … [by various social and causal relations] … This raises the question of precisely how it is determined; and it seems to me that if inflationism is to be believable then the inflationist needs to have some story to tell here… I don't say that the inflationist can't tell a reasonable story about this, only that there is a story to be told,

[48] One might wonder whether the deflationist can avoid this worry by appealing to *interpretability.* According to this thought, the deflationist would insist that the foreign utterance referred in virtue of our *disposition* to *interpret* it by assigning it a particular object. We need not actually interpret or translate it—mere interpretability suffices. (Here, the deflationist is taking a page out of interpretationism.) However, I am doubtful of this move. Interpretations are assignments of objects to expressions, whereas, for the deflationist, word-to-word translation precedes object-to-word interpretation. So for this move to work, the home language must already contain an expression for the intended object. The problem arises when the home language does not have any such expression. In that case, the foreign expression is not even interpretable.

> and perhaps there is room for skepticism about the possibility of telling it adequately. If so, that provides a motivation for deflationism. For the deflationist view is that there is nothing here to explain: it is simply part of the logic of "refers"... that "rabbit" refers to… rabbits and nothing else. (Field 1994a: 260)

> And so we have the deflationist's position. There is no single kind of causal connection that holds between our use of words and their R-referents. There frequently *are* causal connections—even large resembling classes of such connections—and we know, to the extent to which we know our own intellectual history, how these connections were set up. But these connections form too heterogeneous a family to allow anything you might call a general theory of all of them. There is no apparent reason to expect that we can extract from all these connections a simple uniform definition of [reference]. (Leeds 1995: 15)

Let's summarize. The deflationist acknowledges that there are often various causal connections between our words and their referents. They may even acknowledge that these causal connections can serve as the basis for translating one term to another (at least, there's no reason for them not to acknowledge this). However—here is their second gripe with inflationism—they are doubtful that our appeals to these causal connections can be extrapolated into a full-blown *theory* of the reference relation. Apparently, if we're going to explain reference along inflationist lines, then according to Horwich and Field, we need to say what reference is in *non-semantic* terms. And according to Horwich and Leeds, we must also say what the various instances of the reference relation 'have in common'—we must find *unity* amongst all of the various connections between words and referents. Taken together, the inflationist is apparently compelled to find a theory of the form:

(Inflationist Reference / IR) For all x, for all y, x refers to y iff x bears relation R to y

where R is specified in non-semantic terms. (We have encountered this formula before in connection with early Field §1.5.2 and the naturalized content program §1.6.2.)

Well if *that's* the inflationist's burden, then deflationism would start to look quite attractive. Apparently the motivation goes like this. Since the inflationist holds that reference is more than just a disquotation device, they are obliged to give a theory of what reference really is; moreover, they must give their account in non-semantic terms. But this project of *reducing* reference to non-semantic relations is not very likely to succeed (§1.6.2). The deflationist, on the other hand, does not need to give any such substantive theory of reference. They can keep the reference relation trivial and let translation do all of the work. Now, translation, for the deflationist, is not supposed to present the same steep theoretical demands that reference does for the inflationist. This could be for a number of reasons. Perhaps it is because we can develop a use theory of meaning (*à la* Horwich 1998b). Or perhaps we don't need a *general* theory of translation. Maybe translation depends on the goals of the translator (Field 1994a) or maybe we can make do with local explanations for particular translations, without any sweeping generalizations (Leeds 1995). Whatever's the case, the deflationists clearly see themselves as occupying the less burdensome position.

2.5 Historical motive for the second contrast

If this does indeed reveal the deflationist's motivation, then it is fair to ask why they assume that the inflationist is required to give their account of reference in *non*-semantic terms. Why must the alternative to deflationism be a reductive account of reference? To see what is going on here, it is important to place this point into the historical context supplied by the previous chapter. To this end, it is especially instructive to revisit the development of Field's views.

As we saw in §1.5.2, Field spent the early part of his career (before his deflationist days) advocating for an object-based correspondence theory. Moreover, the object-based correspondence theory that he advances in his (1972) has exactly the inflationary structure that was depicted in §2.3. In particular, it sought to explain the truth conditions of sentences as the products of the Tarskian recursive clauses acting on the facts of reference for primitive expressions. In Field's opinion at the time, the remaining philosophical task for this conception of truth was to complete the account by providing a robust theory of reference.

We have also seen in the previous chapter that the early Field placed a rigid constraint on acceptable theories of reference. He required that a correspondence theory must adhere to the formula provided by (IR). The reason for this, which we observed §1.5.2, stemmed from a high-level commitment to physicalism. The motive for analyzing reference in non-semantic terms was to make truth and reference acceptable within a broadly naturalistic worldview.

Finally, we examined in §1.5.2 the attempts to realize this vision that flourished in the decades that followed. This was the naturalized content program. We had also noted that, by the end of the eighties, it was becoming apparent that the reduction of reference to non-semantic relations was unlikely to succeed while upholding all of the guidelines that the program had set for itself. There were several reasons for this, but the most pertinent one has to do with the fact that the selection of reference usually depends on factors that are particular to the term in question. This makes it incredible that there could be any uniform set of conditions, specified non-semantically, that are necessary and jointly-sufficient to determine the referents of all terms. So if this is what is needed for a correspondence theory, then this period of history would show that its prospects are dim.

This is the backdrop in which Field wrote, in 1994, that the reduction of reference to indication is “at best a gleam in the eye of some theorists” (1994a: 255). By the time he had warmed up to deflationism, Field was apparently disillusioned with the agenda that he had set for the inflationary approach.

But despite the setbacks for his former inflationary program, it appears that Field never lost his commitment to physicalism. Moreover, it is easy to reconcile semantics with physicalism if one is willing to accept a deflationary account of reference. All one has to do is define reference like so:

> (RD*) For all x, for all y, x refers to y iff either (x = ‘*a*’ and y = *a*) or (x = ‘*b*’ and y = *b*) or …

where each term in the language is paired with its disquotational referent. In effect, this is just a way of dressing up the deflationary account given by RD into an explicit definition. This account also has the effect of ‘explaining reference away’ since it never mentions reference (or any other semantic notions) on the right-hand-side; so it keeps to the letter of physicalism. The only cost for this reconciliation is that one must be convinced that RD* isn’t missing out on anything important. One must think that there isn’t anything further to reference that calls for explanation. In a word: one must buy into the whole deflationist picture.

2.6 The varieties of explanation

We have thus far uncovered *two* points of contrast between deflationism and traditional inflationism. One key difference is that they employ different orders of explanation for the variety of semantic phenomena. They both recognize the same non-semantic word-world relations, but whereas the inflationist sees these as grounds for reference, the deflationist takes them as local justifications for word-word translations. (We have also found that, judging on this difference alone, inflationism would appear to have the upper hand. The deflationary order seems roundabout and backwards.) The second key difference concerns the *aims* of an explanation of reference. The inflationists are allegedly saddled with explaining reference along the lines of IR, whereas the deflationists are doubtful that this can be carried out. The deflationists, instead, prefer to take the trivial account of reference that ultimately bottoms out in the instances of RD. Finally, we have found a historical motive as to why the inflationist is supposedly burdened with an IR account of reference. The answer, which we saw in Field, stems from the doctrine of physicalism. Given that everything is physical, there are apparently only two options: explain reference away as a device for disquotation, or reduce it along the lines of IR. Since the latter is unpromising, the deflationists chose the former.

In the interest of moving things forward, I propose that we set the issue of physicalism aside for a moment and focus on the explanations of reference. Evidently, much of what is at issue between inflationism and deflationism revolves around the explanations of reference. The deflationists appear to be looking for one kind of explanation and the inflationists only appear capable of offering something else. I would now like to suggest that this whole discussion has hitherto been haunted by a certain ambiguity in what it means to *explain reference*. When accounting for a reference relationship, there are a couple of distinct explanatory demands that could be at play. We would do well to distinguish them.

The distinction that I have in mind is that between *explaining the nature of reference* and giving *selective* explanations of particular reference relationships.[49] (This distinction was briefly mentioned near the end of §1.6.1.) Suppose that our task is to explain a fact of reference—that *'a' refers to b*. What exactly could this demand for explanation impose?

On one (highly intellectualized) reading of the explanatory task, we might hear it as a demand to articulate the *natures* of this fact's constituents. In that case, we would be specifically concerned to identify the *nature* or '*real essence*' of the reference relation. Our task would then be to identify the reference relation with some relation R, which would be taken to *reveal what reference really is*. Doing so would allow one to say that *'a's referring to b* just is *'a's bearing relation R to b*. Since the aim is to characterize reference itself, the account must be general and apply to all instances of reference.

The other kind of explanatory task is much more down-to-earth. For a *selective* explanation of a fact of reference, we would no longer be concerned with what *reference itself* is; instead, our concern would be to explain *why the given term refers to what it does, rather than anything else*. On this way of hearing the demand for an explanation of *'a's referring to b*, we would want to know why it is that '*a*' refers to *b*, rather than *c*, or *d*, or *anything other than b*. We would want to know what underlying states *determine* the fact that *b* was singled out as '*a*'s referent.

This sort of explanation is contrastive (albeit, sometimes implicitly) since its point is to explain why a certain reference fact is actual in contrast to other potential reference facts. The fully articulated structure of such an explanation would thus be:

> '*a*' refers to *b* rather than *c* because '*a*' bears relation *r* to *b* rather than *c*

where *c* is a contrasted object. When a particular relation *r* can play this explanatory role, we say that it *supports selective explanations of reference for 'a'*.

[49] Alternatively, we could characterize it as the distinction between *explaining what reference is* and *explaining what fixes the referent of a term*, or between *the essence of reference* and *the grounds of reference.* The term 'selective explanation' comes from Simchen (2017).

Having this distinction at hand allows us to clarify the deflationist's motivations. From what I've outlined in the previous sections, it is evident that the deflationists are largely concerned with the nature of reference itself. The kind of account that they demand from inflationists—an IR account—is a specific way of characterizing reference's nature: it is a reductive account with a non-semantic analysans. Moreover, it was this kind of explanatory task that occupied the naturalized content program, and the deflationists were motivated by its alleged failure. So, to summarize: the deflationists were looking for a general, reductive account of reference, and since they weren't satisfied with the inflationist attempts, they were driven to their view.

However, now that we have distinguished these two kinds of explanation, we are in a position to offer a rejoinder on the inflationist's behalf. Let's continue to bracket the issue of physicalism and focus solely on the explanations of reference. Let's also grant, for the sake of argument, that a satisfactory IR account of reference is not forthcoming. Would it then follow that deflationism is our only alternative? *No*, because there would still be room for *inflationary selective* explanations of reference.

To see this, consider how we phrased the questions earlier for Inflationary and Deflationary explanations 1&2. There, we wanted to know why my partner's use of '*that*' refers to *o* (the cause of the light) *rather than anything else* (e.g. the distant mountains). We also wanted to know why her term 'Kilimanjaro' refers to Kilimanjaro rather than anything else (e.g. some other mountain, natural object, or artifact). And it is to *these* questions that the inflationary route to explanation appeared to be the superior one. The reason why her '*that*' refers to *o* rather than, say, the distant mountains has plausibly all to do with her referential intentions accompanied by her antecedent causal relations to *o*.

Of course, the deflationists will have their own answer to offer. They will say that my partner's term referred to *o* rather than the distant mountains because '*that*' translates as '*o*' (in my home language) rather than as 'the distant mountains'. They may also claim that this translation is superior precisely because of the causal relations between her term '*that*' and *o* and their resemblance to the causal relations between my term '*o*' and *o*. The deflationists thus have some story to tell about the selective explanations of reference. But when contrasted with inflationism, their appeal to the best translation appears superfluous.

This makes it eminently plausible that the inflationists are on the right track regarding selective explanations of reference.[50] To put this another way, we observed earlier that referring terms tend to bear certain causal-cum-cognitive relations to their referents, and that these relations are explanatorily relevant in some way or another. (My partner's token '*that*' was related by C_1 to *o* and her token 'Kilimanjaro' was related by C_2 to Kilimanjaro.) I'm now claiming that we can plausibly take these causal relations as *supporting selective explanations of their respective facts of reference*.[51] This is just to say that the C_1 relation which '*that*' bears to *o* explains why it refers to *o* as opposed to anything else. Likewise, the causal history traced through C_2 explains why my partner's 'Kilimanjaro' refers to Kilimanjaro rather than any other thing. When we set our sights on this down-to-earth goal for an explanation, these inflationary points seem highly reasonable.

It must also be stressed that these selective explanations do not need to be accompanied by a general account of reference's nature to fulfill their explanatory ends. We don't need to say what reference is *in general* in order to explain why 'Kilimanjaro' refers to Kilimanjaro rather than e.g. Everest—the story provided by Inflationary Explanation 2 will suffice. Our story doesn't need to be universally applicable to every singular term, nor do we need to isolate some non-semantic common factor between each explanation.

In short, we can keep Inflationary Explanations 1&2 and ditch the constraints imposed by IR. Our explanations will still be genuinely *inflationary* since they follow the inflationist order of grounding (particular) reference facts in non-semantic word-world relations. But they are not encumbered by the steep demands of reductive explanation.

Granted, this doesn't go all of the way to answering the deflationist's concerns since we are still setting aside the Fieldian worries about physicalism. In order to address those worries, we need to develop more theory. I will turn to this theory in the next section and chapter. But before I do, I think it is fair to say that this reductionist consideration isn't going to be found as persuasive as it once was. It may be very difficult, or even impossible, to reduce the reference relation to a uniform set of physical conditions. But the same can be said about other kinds of artifacts and sundry goods, and hardly anyone would worry about their physicalist credentials.

[50] No doubt the committed deflationist won't see it this way. They will insist that the prospects for directly grounding reference are dim, so their roundabout explanatory route is a feature, not a defect. To reiterate, I do not claim that the presumptive case I've offered here is intended to win over committed deflationists. Rather, this chapter's main line is intended to undermine a central motivation for deflationism—that reference must be trivial if not susceptible to a general, reductive theory—and set the stage for future debate. (Although once this motivation has been undermined, I do think that deflationism loses much of its appeal.) If this chapter sets up the dialectic right, then the overall case for inflationism, going forward, must lie in the grounding explanations for various kinds of terms. If these explanations are fruitful and serve valuable ends—*and their value cannot be captured within the deflationist picture*—then they present the best positive case for inflationism. Chapter five explains one way in which we can develop this kind of case.

[51] In order for C_1 and C_2 to play this role, they must meet some standards for good explanation. For instance, they should at least support certain counterfactuals like the following: *if my partner's use of 'that' wasn't produced with that intention or perceptual relation to o (codified by C_1) and instead bore an analogous relation to a different object, then it wouldn't have referred to o.*

Hardly anything reduces to anything else *if type-reduction is the standard*. This is a lesson that any inflationist should accept—even those who are sympathetic to physicalism. So if it's true that the naturalized content program failed at its ambitions, then the lesson shouldn't be deflationism; the lesson should be non-reductionism.

2.7 Moderate inflationism

The last section's considerations point us towards a moderate version of inflationism, which I regard to be highly attractive. We have found that the two hallmarks of traditional inflationism—the inflationary order of explanation and an IR account of reference—can be pried apart. This opens up the space for a view that accepts the former while rejecting the latter. The result is a toned-down version of inflationism that isn't an apt target for the deflationists' skepticism towards IR.

The road to moderation begins by endorsing the inflationary structure of explanation for reference, truth, and the rest of the semantic phenomena *for the purpose of selective explanations of semantic endowment.* This means that the inflationary order is employed to answer such questions as: *why does a given sentence 'S' have the truth conditions that it has (as opposed to any other)?* and *why does a given term 'a' refer to what it does (as opposed to something else)?*

The moderate inflationary answer goes as follows. First, the truth conditions of a given sentence 'S' are determined by the referential contributions of 'S's subsentential parts, according to the appropriate rules of semantic composition. (Perhaps these rules were given by Tarski, or perhaps they'll be given by some more sophisticated theory, as noted in §1.5.3.) Secondly, the reference facts for the primitives are explained severally in terms of non-semantic word-world relations. Thus, that a given term '*a*' refers to a particular object is explained directly in terms of some relation borne between '*a*' and that object. We observed in the last section that the C_1 relation can support selective explanations for '*that*' and the C_2 relation can support selective explanations for 'Kilimanjaro'. The idea here is that some such strategy can be employed for each referring term. But the details of the explanation can be particular to the term in question. Contrary to the traditional inflationist, we do not need a *general* theory that abstracts a non-semantic common factor with reference for all other terms.[52] And contrary to deflationism, we do not need our explanations to take a circuitous detour through translation.

This brings us to the second element of moderate inflationism. As advertised, the moderate inflationist has no ambition to give a non-semantic characterization of the nature of

[52] The moderate inflationist is not opposed to seeking semi-general principles to explain and predict reference fixation. The present point is that the viability of this brand of inflationism—as opposed to deflationism—is not tied to the success of discovering such principles.

reference. They have no IR account of reference to offer, and nor do they need one to avoid collapsing into deflationism. For the purpose of selective explanation, all they need are the non-semantic relations that support such explanations of semantic endowment.[53]

This second element of the view might make it sound as if the moderate inflationist is taking too much of an anti-theoretical attitude. Since they don't aim for an IR account, it might sound as if they have no story to tell about the nature of reference itself. But in fact, this doesn't have to be the case. The moderate inflationist isn't precluded from having *any* general theory about reference; their only reservation concerns reduction to non-semantic relations. They deny that reference has a *non-semantic* common factor, but this isn't to say that it has *no* common factor.

So what else can the moderate inflationist say about reference, if not a reductionist theory? I imagine that there is room for several options here. But if you don't mind indulging in a little speculation, I would like to offer my proposal for how the moderate inflationist can understand reference *itself*. This proposal will be highly congruous with the picture given so far.

To organize the discussion, I will state the proposal in this chapter in its briefest form without all of the nuts and bolts. I will say just enough so that I can complete the response to deflationism. Then, in the next chapter, I will fill in the rest of the details.

First, it is widely recognized that, as an alternative to reduction, some qualities (properties or relations) are best characterized by outlining their function (job description, role) within their respective domain. Such qualities are essentially unified by the roles they perform rather than any common underlying condition. This is how I suggest that the moderate inflationist should understand reference.[54] Reference, as a kind of relation, should be characterized, in the first instant, by specifying its essential role.

One advantage of this kind of account is that it allows us to reconceptualize the link between the reference relation and the non-semantic relations at lower levels. Generally speaking, when a quality is essentially characterized by its role, it is said to be realized by the lower-level qualities that perform the role and thereby determine its instantiation. The realizers

[53] In *Naming and Necessity,* Saul Kripke famously expresses doubts that the nature of reference is susceptible to philosophical analysis (1980: 94). He also famously proposes that the reference of a proper name is fixed by a historical chain of communal uses that traces back to the object. To use his example, 'Gödel' refers to Gödel *rather than Schmidt* in virtue of the historical chain that ties the name back to Gödel, rather than Schmidt. Here, Kripke is offering a selective explanation of a particular fact of reference that isn't backed by a reductive account of the reference relation. My proposal so far is thus highly reminiscent of these points from Kripke.

[54] Lynch (2009) employs a similar strategy to develop a pluralist metaphysics of truth. The idea is that truth can be defined by its characteristic role, which is given by various platitudes. He then proposes that any property is a realization of truth provided that it performs truth's essential role for a given domain of discourse.

The view that I am presently advocating is not straightforwardly compatible with Lynch's pluralism. Instead, the moderate inflationist takes truth conditions as always explicable in terms of reference and semantic composition (in the style of early Field) and takes reference to be the multiply-realizable functional kind.

of a functional kind need not have anything in common other than performing a common role. For this reason, a functional kind can be multiply realized.

So then, just what is the essential role of reference? Here is the second part of my proposal. According to my version of moderate inflationism, the essential role of reference is *to explain the dependency of truth value*. When a term '*a*' refers to an object *b*, the essential upshot of this fact is that the sentences containing '*a*' are thereby true or false *depending on the properties of b*. In other words, reference performs the function of binding the truth values of sentences to the properties of things.

If this functionalist analysis is right, then any lower-level relation that can explain the dependency of truth value can play the role of reference. For a relation to perform this function, all it has to do, in effect, is to explain why one particular object, rather than any other, is the one whose properties matter to the truth value of the sentences that contain the term. To put it succinctly, a reference relation is one that explains the *selection* of an object for truth-value dependence.

Given this proposal, we can now see the import of the selective explanations of reference in a new light. We remarked earlier that the inflationist explanations 1 & 2 appear to be the right explanations as to why '*that' refers to o* and *'Kilimanjaro' refers to Kilimanjaro*, respectively. The lower-level relations C_1 and C_2 support selective explanations as to *why* these referents were selected for their respective terms. It follows from this that C_1 explains why the truth value of '*that's* a lighthouse' depends on whether *o* is a lighthouse, and C_2 explains why the truth value of 'Kilimanjaro is the world's tallest free-standing mountain' depends on whether Kilimanjaro is the world's tallest free-standing mountain.

Now suppose that we adopt the functionalist conception of reference that I'm now recommending. We would then regard C_1 and C_2 as *realizations* of the reference relation, owing to the fact that they each play the characteristic role of reference (i.e. supporting selective explanations of truth-value dependence). In that case, the particular fact of *'that' referring to o* is realized by the fact of *'that' bearing C_1* to *o*, and similarly for the fact of *'Kilimanjaro' referring to Kilimanjaro*. But again, this doesn't mean that we can abstract any non-semantic common factor between the lower-level facts for the purpose of a reductive account. When it comes to functionalist metaphysics generally, there may be no underlying unity between the multiple realizations of a functional kind, other than that they perform a common role. Nonetheless, there is still another kind of unity to a functional kind. In this case, each realization of reference will have this much in common: they each serve to explain the selection of objects which make for the truth-conditional contributions of words.

Earlier, we observed that the deflationists were motivated by the presumption that their position is less theoretically burdensome. According to this thought, all of the deflationist's burden gets offloaded onto their account of translation—for instance, their explanation as to why ‘*that*’ translates as ‘*o*’. And although they may appeal to word-world causal relations to explain this translation, unlike the traditional inflationist, they make no attempt to devise a systematic theory of the non-semantic connections between words and their referents. Instead, they can justify their translations on a case-by-case basis. For this reason, deflationism is supposed to be less demanding.

But now that moderate inflationism is on the table, the dialectical situation has been reshaped. Much like the deflationist, the functionalist is not encumbered by the exigencies of providing a systematic account of word-world connections in non-semantic terms. Similar to the deflationists, they too may operate in a piecemeal fashion; they can offer their selective explanations of reference for each term taken individually. So it is doubtful that the deflationist has an advantage here.

The final point to make is that we have now also reached a proper way to respond to the Fieldian worries about physicalism. Recall from earlier that the deflationists were motivated by the alleged difficulty in reconciling inflationary semantics with the physicalist worldview. Well, if the functionalist version of moderate inflationism is viable, then we can lay these physicalist worries to rest. That is because a functional property or relation is (generally regarded to be) compatible with (non-reductive) physicalism when all of its realizations are. So if all of the lower-level relations that play the role of reference are compatible with physicalism, then so are reference and truth. But if not, then not. The point is, the Fieldian worries about physicalism present no immediate problems for this view.

2.8 Conclusion

Let me summarize the main thread of this chapter. As I’ve been telling it, the main point of contention between inflationists and deflationists was never so much about *truth*, but rather, about *what it takes to explain reference.* Inflationists take (non-semantic) word-world relations to explain referent selection, whereas the deflationists wanted something more ambitious: a general reduction of reference to non-semantic relations. Failing that, the deflationists claimed that the explanation of reference *and truth* ought to be kept trivial. In my view, an inflationist ought to reject the deflationist’s demand for reduction and keep their non-trivial explanations for how particular referents get selected. Doing so paves the way for a moderate version of inflationism. Such a view would retain the idea that reference facts are explained on an individual basis by word-world relations, and are unified together by their shared role in explaining truth.

To situate this conclusion within the broader context of my project, recall that in chapter one it was argued that a modern object-based correspondence theory requires a theory of both semantics *and* metasemantics. In this chapter, we have now seen how the over-extended ambitions for the metasemantical component gave way to deflationary skepticism towards this entire conception of truth. (This appears to be why Field abandoned his early vision for the theory.) However, we have also seen that deflationism was motivated by historically contingent reasons. They saw their opponents as tied to a rigid reductionist program which, in hindsight, had little chance to succeed.

By dividing up the tasks for the metasemantical part of the theory and keeping them at a reasonable level, we are able to revive the spirit of the object-based correspondence theory (although not quite in the same form as Field 1972) and keep it immune to the deflationist's objections. To be sure, the theory still needs a general orientation for selective explanations of reference—i.e. productivism—and an account of the reference relation itself. But I claim that these are relatively independent projects. In this chapter, I have only given a brief proposal for how the view can deliver on the second project of articulating a theory of reference. But the functionalist account that I hinted at is still short on the details. In the next chapter, we turn our attention to bringing this account up to scale.

Chapter 3: A Pluralist Theory of Reference

3.1 Introduction

Contemporary thinking about truth has recently experienced a trend towards *pluralism* (Wright 1992; Lynch 2009; Sher 2015, 2016; Edwards 2018). The common idea is that truth is multiply realizable, and is realized by different properties in different domains. According to this view, truth may be realized by (some kind of) correspondence for thought and talk about mid-sized ordinary objects, while also being realized by (some kind of) coherence property for ethics or mathematics. Lynch (2009) develops this idea by employing an analogy to functionalism in the philosophy of mind. His idea is that truth ought to be characterized, in the first instant, by specifying its essential function, invoking various 'platitudes' about truth, and then allowing for it to be realized by any property that fulfils this function throughout various domains.

One drawback of the pluralist theories of truth is that they do not sit well with a certain, compelling view that sees truth as a matter of accurate representation. This alternative approach takes advantage of the progress made in semantic theory and the theory of reference to cash out truth conditions as the products of semantic composition and primitive reference relations. Glanzberg (2015) calls this the *modern correspondence theory*, and it is found in the works of (early) Field, Kripke, Devitt, Kaplan and Fodor. Davidson (1977) also dubs it the 'building block' conception of truth since it seeks to build the conditions for truth on top of the facts of reference (although he uses this as a term of derision). At the level of truth, this approach offers a uniform analysis: truth is *always* a matter of truth-conditional satisfaction. It is thus doubtful that this picture can be reconciled with truth pluralism.

Nonetheless, it is possible to borrow some of the insights from the pluralist theories and adapt them to the needs of the modern correspondence theory. The problem for modern correspondence is that *reference* appears to resist any uniform analysis. Davidson, for example, complains that reference has no 'empirical content' (1977). However, if instead of applying the metaphysics of pluralism to *truth,* we apply it to the relation of *reference,* then this problem can be sidestepped. Indeed, there are even some recent precedents for making this move; see Palmira (2018) and Moore (2022).

In what follows, I would like to present a theory that takes the best of both worlds: a pluralistic metaphysics for primitive reference relations and a semantic compositional view of truth conditions. Let's call the former component the *pluralist theory of reference*. The key ideas of the pluralist theory are: (i) on an individual basis, it is possible to provide reference-fixing stories for each term and referent; (ii) the stories need not have much in common other than that they play a distinct role for the purpose of semantics; (iii) the role for semantics is what unifies each instance of reference.

3.2 Selective explanations and the nature of reference

One hallmark of this theory is that it decouples several of the distinct explanatory aims that the traditional theories tend to run together. Before we turn to the theory, it is important to distinguish them. In the previous chapter, we characterized two such aims: the task of giving *selective explanations for specific reference relationships* and the task of *characterizing reference itself*.

To get a sense for the first kind of explanation, it is best to consider an example. Suppose that we are walking on a beach and we see a dim light in the distance. I point to it and say 'that's a lighthouse'. We now ask: why does my token 'that' refer (on this occasion) to the object—call it *X*—that is the cause of the light, rather than anything else? Presumably, this has an answer: it is because I am intending and gesturing to bring your attention towards *X*, which is the object that I am perceptually focused on, and not anything else.

In offering this answer, notice that I am not thereby attempting to say what reference *is*, or to state *general* conditions for the fixation of reference. I am merely bringing up salient facts about my speech in order to explain why the particular referent was selected for this term. The aim for this explanation is thus extremely modest; it is a far cry away from the ambitions of the traditional theories of reference. Not only that, but as long as we limit our sights to this circumscribed task, this explanation can hardly be objectionable.

What I have offered is a selective explanation of reference. This kind of explanation is fairly undemanding, and there are several features that make it so. For one, observe that we are only concerned with a specific, given referring term—namely, my token 'that' in the utterance 'that is a lighthouse'. We were not concerned with the referential features of any other term. Also notice that we are concerned with a certain 'why'-question. The purpose is to explain why the term refers to its actual referent, as opposed to other potential referents. Each of these features is crucial to the distinctive aim of a selective explanation of reference; the aim is to explain why the referent for the term was selected in contrast to other objects.

Generally speaking, a selective explanation for a given referring term *t*, which refers to object *o*, will have the following form:

> (1) *t* refers to *o* rather than anything else because there is a relation *r* such that *t* bears *r* to *o* and *t* does not bear *r* to anything else.

To *answer* the task of a selective explanandum, we appeal to some relation *r* that obtains between the term and its referent and does not obtain between the term and any other object. In the story given above, it suffices to point out that I am *intending* to direct your attention towards *X* while perceiving *X* and not towards anything else (e.g. the mountains on the distant shoreline). The relation defined by my intention and perception serves this explanatory role. When a relation *r* plays this role, let's say that it *supports* the selective explanation of reference. I will say more about the relations that are suitable for this role in the next section, but for now, the chief purpose is to clarify the distinction between the two kinds of explanation.

Now let's oppose this with a very different sort of task that one might expect from a theory of reference. Traditionally, most theories in this literature aspired to do much more than offering mere selective explanations. Instead, their primary aim was to say *what the nature of reference (itself) is.* Not only that, but they aimed to do so by uncovering a general relation that was to be identified with reference. The final form of such a theory would be:

(2) For all x, for all y, x refers to y if, and only if, x bears R to y.

Such a theory would allow one to say that *t's referring to o* just is *t's bearing relation R to o.* The aim would be to elucidate the reference relation by analyzing it in terms of *R*.

Perhaps it would be helpful to think of this distinction in terms of grounding and essence. For a selective explanation, we seek to explain *why* a given reference fact holds. That is, we seek to articulate the *grounds* for a fact of reference—i.e. the 'lower-level' facts that determine its obtaining at the exclusion of other potential reference facts. For the purposes of a grounding claim, the nature of reference itself is presumed to be unproblematic. On the other hand, for an account of the nature of reference, the reference relation itself is the sole issue, irrespective of which terms bear it to which objects. The point is to say what reference *is*—what its *essence* is. The determination of particular reference facts is not the main concern.

The history of the last half-century has known numerous examples of theories that were intent on delivering accounts of the nature of reference. The most well-known ones (for mental representations pertaining to kind terms) include:[55]

- The theory that reference is a certain kind of causation (Stampe 1977).
- The theory that reference is a certain kind of informational relation (i.e. lawlike correlation) (Dretske 1981, 1988; Fodor 1987, 1990).
- The theory that reference is a certain kind of teleological relation, e.g. what a representation functions to indicate (Millikan 1984; Papineau 1984, 1993; Neander 2017;

[55] To be fair, these theories were often not intended to describe reference for every kind of representation—e.g. they were not intended to capture reference to numbers.

Neander & Schulte 2022).

Each one of these basic ideas has been developed into enormously sophisticated accounts. However, rather than exploring the details, I mention them only to present them as rivals to the approach taken in this chapter.

The traditional approach to the theory of reference typically operated with further ambitions. For one, there was a widely-shared commitment to metaphysical naturalism which placed steep constraints on how reference ought to be analyzed. This manifested in the goal of analyzing reference in *non*-semantic terms, since this was thought to be imperative for reconciling semantics with the naturalistic worldview. Secondly, the theories also endeavoured to give *general* (i.e. necessary and jointly-sufficient) conditions for reference fixation. These were meant to explain why *any arbitrary* term refers to its referent. Should a theory deliver such conditions, it would then be able to explain the selection of reference for specific terms by subsumption under general law. So the traditional theories do share the aim of selective explanation. However, unlike the example offered earlier, they approached them from a place of general, overarching theory.

By trying to do all of these things at once, the traditional theories had their work cut out for them. Indeed, if all of these goals are required for success, then success may be unattainable. But I won't press this point here. (Others have already done so; e.g. Loewer 1987; Boghossian 1989.) Instead, I would like to make the point that we don't need to impose upon a theory all of these goals as joint constraints.

Here is an observation that I hope to be fairly uncontroversial: given the limited aims for selective explanations, they do not require an account of reference's nature to meet their explanatory ends. In saying this, all that I mean is that the example of a selective explanation that was offered earlier is sufficient on its own. We do not need to say what reference is *in general* in order to explain why my token of 'that' refers to *X* rather than anything else (e.g. the mountains on the shoreline). Nor do we need to say what the necessary-and-jointly-sufficient conditions for reference-fixation are for arbitrary terms. The appeal to my intentions and perceptual state is enough to explain why my token refers to nothing other than X.

If this is right, then the theory of reference is permitted to begin with a particularist methodology. We are permitted to take the selective explanations for each particular term as a given, and then build our theory on top of them, like scaffolding overlaying a fixed base.

3.3 The relations that support selective explanations

In the example given above, we were chiefly concerned with a demonstrative expression in public language. For this particular expression, the selective explanation was best served by appealing to speaker intentions and perceptual states. For other expressions of public language, we might appeal to a litany of other factors, depending on what kind of term we're concerned with. A survey of other factors that loom large in the literature must include:

- Acts of ostension, especially for natural kind terms (*à la* Putnam 1975).
- Intentions to defer to whomever one learned a name from (*à la* Kripke 1980).
- The exploitation of linguistic conventions to settle the reference for context-sensitive expressions (*à la* Kaplan 1989b).

And no doubt there are others. For the purposes of a pluralist account, there is no need to be a completist about the potentially relevant factors.

As we can see, the selective explanations for public language expressions may take the intentional features of our mental states for granted. But in addition to public language, we must also consider the selective explanations for the intentional mental states. When I uttered 'that's a lighthouse', I also *believed* <that's a lighthouse>. We can also ask why my belief is about *X* rather than anything else. A natural answer is that my belief was formed in the aftermath of perceiving *X*. Moreover, my perceptual state was about *X* because it was formed as the result of a causal process that traces back to *X*. Among other things, this causal process must include the fact that my perceptual focus was directed towards *X*. We can thus say that my belief is about *X* because it was produced under such-and-such circumstances, mentioning the direction of perceptual focus towards *X*.

When it comes to other intentional states, I take it that there will always be *some such* relation that can explain the selection of content. If we ask why the mental state *M* is about (i.e. refers to) *o* rather than anything else, then there will always be *some* features about *M*'s production that tie it back to *o* that can support the explanation in question.[56] We may not always be in a position to articulate what they are, but that poses no problem to this thesis since the concern is metaphysical explanation, not epistemology.

When we trace back the production of a referring item (a public language expression or an intentional mental state) through its causal history to its referent, we can thereby trace

[56] By emphasizing the features that surround the *production* of the term or state in question, I am hereby aligning this view with the 'productivist' orientation in metasemantics (see Simchen 2017), as opposed to metasemantic interpretationism. For present purposes, all that means is that a selective explanation for a specific reference relation is best supported by features surrounding the production of the referring item, and not by appealing to an interpretive theory of the subject's overall linguistic and cognitive behaviour.

numerous relations that are not themselves *semantic.* By this, I mean that the description of these relations need not mention *reference, aboutness, satisfaction,* or *truth.* But even though these relations are not themselves semantic, I claim that, within a specific explanatory context, they may be apt to support selective explanations of reference. If, say, the relation describes my visual system as focused on *X* rather than (for instance) the distant mountains at the time of uptake into my cognition, then that would be part of an explanation as to why my subsequent 'that' tokens are about *X* rather than the distant mountains.[57]

In order to build the subsequent theory, it is important to generalize from these examples. To this end, we must advance a thesis that says, in effect, that selective explanations of reference are always available for any referring term. The particular version that I propose is this.

The Semantic Grounding Thesis (SGT). For all terms *t*, for all objects *o*, *t* refers to *o* if, and only if, there is a nonsemantic *r* such that (i) *t* bears *r* to *o uniquely,* and (ii) for all other objects *c ≠ o, t* bearing *r* to *o* rather than *c* explains why *t* refers to *o* rather than *c.*[58]

This thesis entails that selective explanations will always be supported by a non-semantic relation between term and referent. In short, the facts of reference are grounded in the non-semantic facts.

A few clarifications about SGT are in order. First, what relations '*r*' are within the scope of this thesis? So far, all that has been said about them is that they are non-semantic, pertain to term production, and support selective explanations. One might wonder what else must be true of these relations for them to fulfil these roles.

It is at this point that the *pluralist* aspect of the theory enters the picture. A theory of reference (or a theory of content) must have further explanatory aims and constraints to narrow the range of acceptable explanations. However, unlike traditional theories, a *pluralist* theory does not impose these additional constraints from the top down. Instead, it allows for a variety of local explanatory contexts, where the explanations of reference may be tailored to meet whatever explananda are appropriate to those contexts.

I'll give a few examples. A theory of representation in cognitive science is geared towards explaining the content of (typically sub-personal) mental representations for the mid-

[57] Against this suggestion, some philosophers will contend that the argument of Kripke (1982) shows that the non-semantic facts are insufficient to ground the normative element of the semantic facts. For this reason, it is claimed that no non-semantic relation can explain why by '+' we mean *plus* rather than *quus.* To respond to this worry, I appeal to Boghossian's (1989) response to Kripke (1982). Boghossian contends that Kripke's (1982) overlooks the possibility of a non-reductionist, robust realist account of meaning (or reference). In effect, the account of this chapter is a version of such a theory.

[58] The uniqueness clause in (i) assumes that *t* determinately refers to only one object. This is an idealization. To account for the phenomena of semantic indeterminacy, adjustments will have to be made. I will not consider them here.

sized objects that organisms must successfully deal with to survive and thrive. As such, its central practice is to explain successful behaviour by appealing to successful representation, and a theory of content must underwrite this practice (Shea 2018: 10). Now, in contrast, when we consider *personal-level* intentional states (e.g. beliefs and desires) about mid-sized sundry goods, we might find different explanatory aims appropriate. There is a longstanding practice within philosophy (going back at least to Russell 1997, 2007) of subjecting the theory of reference to epistemic constraints. Within this explanatory setting, an account of reference may also be required to explain how the subject achieves *epistemic* access to referents (see Dickie 2015 for a recent example). Thirdly, consider Putnam (1975)'s causal account of reference to natural kinds and substances. Here, the concern is scientific representation, and Putnam's specific aim is to defend the realist's conception of science. An account of reference in this setting will endeavour to explain how our scientific theories manage to represent a realm of objective substances and kinds.

Each of the above examples has a common aim to explain reference to *natural* objects, substances, and kinds. This, in itself, imposes a constraint on the relations that serve these explanations: they must be suitable for this kind of explanation. It's not easy to say exactly what this constraint amounts to, but it is widely recognized that some qualities have explanatory power, whereas others don't. The real explanatory relations form an elite minority. As a first pass, the relations that are (in some sense) *natural*—that figure into the natural sciences, that delineate the joints of nature—are prime candidates. This includes causal relations, physical relations, perceptual relations, and the facts of inner cognitive processing.

However, as Chomsky (1995) famously argued, reference relations *in general* cannot be scientific because *referents* are (typically) unscientific; most of the things we think and talk about aren't carved out by nature—they are delineated by our own human interests. The pluralist approach is well-suited to handle this observation. According to the pluralist theory, the relations that explain non-scientific reference (e.g. reference to socially-constructed kinds) need not answer to the same explananda as scientific reference. They can instead answer to the explananda that are appropriate to their own explanatory settings. When it comes to social kind terms, we should expect that elements from our social reality—i.e. our culture (including conventions) and our personal interests—will be relevant to explaining reference fixation.

The point is, the pluralist theory, by itself, leaves it fairly wide open what a selective explanation of reference must look like. In this way, it respects the autonomy of the special sciences and the various philosophical disciplines. They may each pursue their metasemantic questions as they see fit for their own domain. I will return to this point later and explain why it is a virtue.

The second thing to notice about SGT is that it is consistent with the grounding relation varying from term to term. Indeed, a plurality of explanatory ends will likely lead to a plurality of explanatory means. What grounds the fact that *'that' refers to X* may be quite different from what grounds the fact that *'Kilimanjaro' refers to Kilimanjaro,* or the facts of reference for other perceptual demonstratives or numerals or social kind terms. This is a desirable feature. We do not want to make the unrealistic claim that reference affords a uniform analysis of all of the many ways we use words to represent things. For all that SGT says, there may be little in common in the explanation from one term to another. Reference may have a plurality of grounds. However, also notice that SGT is consistent with the plausible idea that there are similarities, or semi-general laws, that pertain to terms of a common kind (e.g. perceptual demonstratives, numerals, social kind terms, etc.).

The last thing to observe is that SGT requires that selective explanation is *contrast invariant.* Each fact of reference is grounded in (at least) *one* non-semantic relation that supports selective explanations *relative to any contrast*. SGT doesn't allow the grounds for a fact of reference to vary depending on which contrast is mentioned.

Is this contrast-invariance justified? Or might it be that the grounding relations involved in metasemantics are irrevocably contrastive, and different contrasts call for different grounds? Schaffer (2012) argues that metaphysical grounding is generally contrastive. Perhaps there is space for a view that sees reference fixation—a species of metaphysical grounding—as inherently contrast-sensitive. I have not myself ever seen this question posed in the reference literature. However, it is beyond this chapter to give it our full consideration.[59]

There is a rough, impressionistic motivation for preferring SGT over a contrast-sensitive view of semantic grounding: it is better aligned with a *realistic* outlook towards the semantic facts. Admittedly, this is another classification that is hard to make precise. But if we are going to understand the semantic facts as emerging from real features of the world, then we should expect them to be susceptible to explanations, predictions, and laws. This is consistent with the metaphysics of pluralism, but it is hard to see how it would be possible if semantic explanations varied according to which contrast is mentioned in the explanandum. Now, I intend for the pluralist theory developed here to be in harmony with this realistic attitude. Hence, I will submit SGT as the grounding thesis upon which it will rest.

It matters a great deal to my pluralist theory whether the semantic grounding thesis is true. Yet obviously, it would be poor form to justify it solely on the basis of one example. So we must now confront the question of its truth, and we must say more in its favour besides the

[59] To argue for such a view, one would have to find an example where (i) *t* refers to *o*, (ii) *t* does not refer to *a* or *b* ($a \neq b$), (iii) that *t* refers to *o* rather than *a* is explained by the fact that *t* bears r_1 to *o* rather than *a*, (iv) that *t* refers to *o* rather than *b* is explained by the fact that *t* bears r_2 to *o* rather than *b*, (v) $r_1 \neq r_2$, (vi) that *t* refers to *o* rather than *a* cannot be explained in terms of r_2, (vi) that *t* refers to *o* rather than *b* cannot be explained in terms of r_1, and finally (vii) that *t* refers to *o* rather than *a* and that *t* refers to *o* rather than *b* cannot both be explained by another relation *r* that *t* bears to *o*. Not only that, but *t* must refer to *o* unambiguously throughout the example; the variation behaviour cannot be chalked up to semantic ambiguity. At present, I do not know of any convincing counterexamples to SGT that fit this formula.

example. To this end, let’s consider what it would take for the semantic grounding thesis to be false.

Once unpacked, the semantic grounding thesis makes several distinct assertions. For one, it entails a supervenience thesis. This amounts to the claim that a difference in reference locally supervenes on the non-semantic relations between words and objects. In other words, there’s no difference in reference without a difference in non-semantic relations. Secondly, it adds that each term and referent has a non-semantic subvening relation that *explains* (or *grounds*) the difference in reference. Lastly, it entails that the grounding relation is contrast-invariant. Since we have already discussed the last component, let’s turn our attention to the other two. If one of these were false, then either one of two things would have to happen: either there’s a difference in reference that doesn’t supervene on the non-semantic relations, or there’s a difference in reference that isn’t explicable by the subvening relations.

When we put the semantic grounding thesis this way, we see that it amounts to nothing more than a very weak thesis about how the semantic level relates to the non-semantic levels. We might put it sloganistically as the claim that semantic facts are not fundamental or brute; they are grounded in non-semantic facts. As a high-level principle about how semantic facts fit into the world, this seems to be a safe working assumption. On the other hand, to deny SGT, one must claim that there are some semantic differences that are inexplicable *in principle.* Some differences in reference obtain, and there is nothing more fundamental that could explain them. I regard this to be highly incredible.

Nonetheless, if there are potential counterexamples, then they need to be given a hearing. To my knowledge, the most significant challenge to SGT comes from Robert Brandom and concerns complex numbers. According to the theory of complex numbers, each negative number —for instance, -1—will have two square roots. Suppose that we define ‘*i*’ as one of the square roots of -1. In that case, we must ask: *which one?* This question turns out to be problematic for the present view because the two square roots of -1 share all of the same structural properties. There is a perfect automorphism of the complex plane that maps each complex number to its complex conjugate and preserves all of the relevant structure. Moreover, since the square roots of -1 are numerical objects, their features are exhausted by their position in the system of complex numbers, so they cannot be distinguished by extra-structural intrinsic qualities. For this reason, it is difficult to see how there could be any non-semantic facts about our usage of ‘*i*’ that explain its reference to one square root of -1 rather than the other. We may presume that ‘*i*’ refers to exactly one of the two square roots of -1, but it is otherwise inexplicable as to why it refers to the one it does.

This does indeed look awkward for the semantic grounding thesis. But does it present a decisive refutation of it? Should we abandon all hope this early on? I believe that such despair would be premature. After all, counterexamples involving mathematical objects launch their attack from a position of shaky philosophical ground. Brandom himself is explicit that the problem only arises for the theory of reference from within the platonist setting (1996: 295). So to get this objection started, we have to have already taken a number of controversial steps into the metaphysics of abstracta. For this reason, I do not think it would be too irresponsible to raise this objection as a curiosity and then set it aside for the purposes of this chapter. From now on, I will assume that SGT is true.

3.4 The role of reference

The claims of the previous section can now be brought together into the service of a pluralist metaphysics of reference. My plan is to model this pluralist theory on an account of truth given by Michael Lynch (2009). In this book, Lynch develops a functionalist metaphysics of truth by analogy to psychological functionalism in the philosophy of mind. He thus proposes that truth should be characterized by its function and that it is realized by a plurality of diverse lower-level properties.

Generally speaking, there are three features that distinguish a functionalist account of some property/relation *F*. First, the theory will claim that *F* is characterized by its causal or explanatory role within its respective domain; secondly, *F* is said to be realized by various 'lower-level' properties, provided that they fulfill the requisite role; and thirdly, the properties/relations that realize *F* need not have anything else in common—*F* may be *multiply realizable.*

The central idea that I would like to explore is that *reference,* as a general relationship, can be characterized along functionalist lines. I thus propose to characterize the reference relation by its role and then allow for it to be realized by any lower-level relation that plays its role.

The first step to developing such a theory is to identify the role of reference. To this end, it is key to remember that for a functionalist account, it is permissible to specify the role of a functional kind from *within* its own domain. In this case, we are dealing with a relation that is squarely at home in semantics. We are thus permitted to identify reference's role by tracing its relations to other semantic phenomena, such as truth and meaning. This is more or less how I plan to proceed. My main claim of this section, then, is that the role of reference is given by its position *within* semantics.

Here I am thinking of semantics in its usual truth-conditional form. In this form, semantic theory is in the business of explaining how the truth conditions of whole sentences are determined by the semantic contributions of their smallest significant parts. To give the simplest examples, a semantic theory is meant to capture such pre-theoretic facts as:

- 'Fa' is true if, and only if, 'a's referent instantiates the property expressed by 'F'
- 'aRb' is true if, and only if, 'a's referent bears the relation expressed by 'R' to 'b's referent.

And so on. Within the formal setting, a semantic theory will posit some mechanism of composition (e.g. functional application within the Fregean paradigm) that will determine the kind of mathematical objects that may serve to represent the semantic contributions of each kind of expression. All of this is done in the service of providing a formal representation of how sentential truth depends on the significance of subsentential expressions, including the referents of singular terms, given the theory's modelling assumptions.

However, it is also important to see that a compositional semantic theory, as such, is not in the business of saying *why* or *how* any given term refers to what it does. A theory of semantic composition will tell you why *'snow is white' is true if, and only if, snow is white* given that *'snow' refers to snow* and *'is white' applies to white things*. But it does not offer any deep explanation of the latter two facts. For the purposes of semantic theory, those facts are taken as given. (This is why questions about reference fixation are sometimes called '*pre*-semantic'.) We might put this point by saying that semantics is really only aimed at two things: to uncover the compositional structure of truth, and to offer selective explanations of sentential truth conditions *given* the facts of subsentential meaning.

The role of reference, then, is to provide the basic inputs for semantic composition. Its role is to assign objects to terms so that the semantic machine can do its work in generating truth conditions. In doing so, each reference relation will thereby explain why the truth values of sentences depend on the properties of particular objects. For example, the sentence 'that is a lighthouse' is true or false depending on whether *X* is a lighthouse; however, its truth value does not depend on whether the mountains on the shoreline are lighthouses. But why does its truth value depend on the properties of *X* rather than the mountains? Answer: because the token 'that' refers to *X* rather than the mountains. Simply put, reference performs the function of binding the truth values of sentences to the properties of things.

It is possible to describe this function more explicitly, and if you don't mind the extra pedantry, it will pay off in the next section. The function of reference is captured by the principle:

> (3) that *t* refers to *o* explains that for all sentences *S* in which *t* partakes, *S* is true if, and only if, *o* bears the property (or structure of properties) expressed by the remaining constituents of *S*.

(The usual provisions must be taken to deal with context sensitive expressions. Other provisions may have to be made for opaque contexts.)

It must also be emphasized that this explanatory function is chiefly a matter of selection. The function of reference is to *select* an object, from among all objects, for truth-value dependence. There is thus an implicit contrastive element to this principle. To make it explicit, we should say:

> (4) that *t* refers to *o*, rather than anything else, explains the fact that *S*'s truth value depends on *o* bearing the relevant property (or structure of properties, expressed by the remaining constituents of *S*), rather than anything else bearing that property (or structure of properties), for any sentence *S* in which *t* partakes.

Either of these will suffice to exhaust the function of reference within semantics.

It is a striking fact about reference that the best way to say what it *is* is to outline what it *does* in relation to other semantic properties, such as *truth.* This makes it highly plausible that a functionalist account of reference is the right way to go.

3.5 The realizers of reference

The foregoing section gives us the first ingredient for a pluralist theory of reference: a description of reference's function. But it still remains to identify the lower-level relations that facilitate this function. Since reference is a semantic notion, the 'lower-level', in this context, includes anything that is *non-semantic*. So the task is to find the non-semantic relations that play reference's role. Thankfully, since we have the semantic grounding thesis, we are well-positioned to do this.

According to my claim in the last section, the function of reference within semantics is to assign objects to names for the purpose of generating the truth conditions of sentences. In other words, it explains *which* objects matter to the truth values of sentences. Now, according to SGT, each referring term will have *some* non-semantic relation that explains the selection of referent for that term. If we put SGT and (4) together, we get the result that:

> (5) For all terms *t*, for all objects *o*, if *t* refers to *o*, then there is a nonsemantic *r* such that (i) t bears *r* to *o* uniquely, and (ii) *t* bearing *r* to *o* explains why *S is true if, and only if, o bears the relevant property (or structure of properties, expressed by the other constituents of S), for any sentence S in which a partakes.*

And all that this says is that each referring term enjoys some non-semantic relation to its referent that fulfills the characteristic role of reference. So SGT combined with my claim about the role of reference entails that reference is always realized by some non-semantic relation.

Now that the pieces have all come together, it is worth reiterating how this picture unfolded from where we began. It all started with the observation that, for each referring term, it is not too difficult to say why it refers to one thing rather than any other thing, provided that is all that we aim to do. Then, from the availability of these explanations, we advanced the semantic grounding thesis, which guarantees that each term and referent is joined by some non-semantic relation that can support selective explanations. Following this, I claimed that the job of a reference relation is to explain why the truth values of certain sentences (the ones containing the term) will depend on the properties of a particular object. Now, since the relations that explain the selection of referent can effectively do just that, we may thereby conclude that they realize reference as well.

Earlier I noted that the grounding relations guaranteed by SGT may appeal to factors that pertain specifically to the term and object in question. For this reason, it follows that the realizations of reference may be, for all that this theory cares, a highly heterogeneous and motley assortment. But this isn't really a problem for the theory. Functionalism allows for the realizers of a functional kind to be diverse and disunified at their own level: they need not have anything in common other than sharing a common role. In this case, the realizers of reference are still unified by their shared semantic function of explaining truth.

Since I am advocating for a functionalist theory, I'm obliged to say a word about the relation between the functional relation—*reference*—and its realizers—the non-semantic explanation-supporting relations between words and things. In this connection, I must also say a word about the distinction between realizer functionalism and role functionalism. Briefly put, the theory needs to take a stand on whether the functional property/relation (reference) is numerically identical to its realizers in each case, or if it is always distinct from them. The former leads to a kind of reductionism, whereas the latter leads to a kind of non-reductionism. And if reference is distinct from its realizers, then we also need some account of what else the relation between them could be.

On these questions, I would like to take the theory from Lynch (2009) as a model. Unlike the account developed here, Lynch is primarily concerned with the property of truth. Nonetheless, his account of the relation between the role-property (in his case, truth) and its realizers can be adapted to suit the pluralist account of reference.

As a version of role functionalism, Lynch's account takes the functionally-characterized property to be distinct from each of its realizers, but unlike other versions of 'role functionalism', he does not construe the functional property as a mere higher-order property.[60] Instead, we can gloss Lynch's account as claiming that functionally-characterized qualities are *abstractions* from their realizers.

Here is how to make this more precise. First, for each property and relation, there is a distinction between its nominal features and its essential features. The essential features are what make a property/relation what it is, whereas the nominal features are the ones we use to pick it out (they need not be essential). To use an example that we've encountered before, the traditional theories of reference would have nominally picked out the reference relation by its role in semantics and then sought to describe further, underlying *essential* non-semantic features that each reference relation has in common.

For Lynch, what distinguishes a property/relation as a *functional* property/relation is that its essential features are (more or less) exhausted by its role.[61] Adjusting this claim to the present case, our claim that reference is individuated by its function is just the claim that reference is essentially the relation that plays the role captured by the principles (3) and (4). Reference *just is* the relation that ties objects to words for the purpose of determining truth conditions, and that's all there is to reference *as such.*

If this is right, then according to Lynchian metaphysical principles, a functional property/relation will have a fairly thin set of essential features. This means that other properties/relations may share all (or nearly all—mind the footnote) of the features that make a functional property/relation what it is, plus more besides. The essential features of a functional property/relation may form a (near) proper subset of the features of another distinct property/relation. When this occurs, we say, in Lynch's terminology, that the lower-level property/relation 'manifests' the functional property/relation. Moreover, since the essential properties of a functional property are

[60] By claiming that the functional property/relation is distinct from its realizers, we avoid the undesirable consequence, endemic to realizer functionalism, that there is no unified, overarching functionally-defined property/relation. If that were our claim about reference, then we'd have to admit that the relations instantiated by the pairs <'Kilimanjaro', Kilimanjaro> and <'Everest', Everest> are similar in name only, and that there is no common *reference* relation shared by the two pairs. Since this appears to be the wrong result, I opt for role functionalism instead.

In the philosophy of mind, it is common to see 'role functionalism' construed as the claim that a functional quality *F* is a higher-order quality—namely, the quality *of instantiating some realizer of the role R.* See McLaughlin (2006) for a criticism of this view. If this were our account, then reference would be the higher-order relation instantiated by a pair <*a*, *b*> if, and only if, the pair instantiates some other non-semantic relation that plays the role of reference. This account is worrisome because it implies that the role quality (reference) doesn't perform the role itself; it is the lower-order realizations that perform the important role. The functional quality serves as a mere stamp of approval that the job has been done elsewhere. For this reason, I do not endorse this particular brand of role functionalism.

[61] Why did I say 'more or less'? Because, besides its role features, a functional quality *F* will also have a few other essential features that must be ignored to define Lynch's 'manifestation' relation. Besides its role-related features, reference also has the essential features of being *a semantic relation*, *being identical to itself,* and (if I'm right) *being a functional relation.* But let's ignore these non-role-related essential features. It's enough that *almost* all of the interesting essential features of a functional quality have to do with its role. Nevertheless, for this reason, the reader will have to mind the hedging that ensues for the remainder of this section.

a matter of its role, it follows that any property that *manifests* a functional property is *ipso facto* facilitating its role. For Lynch, this is the connection between a functional property/relation and its realizers. In this sense, the functional properties/relations can be seen as abstractions from their realizers; their essential features are (for the most part) a matter of subtracting the specificity of their realizers.

Taking Lynch's theory as our model, here is what I would like to say about reference. Reference is the relation borne between a term *t* and object *o* when and only when *t* contributes *o* to the truth conditions of all sentences in which *t* partakes. That is essentially what reference is, and there is nothing much more to say about reference *as such*. However, in each case, there will also be a non-semantic *r* borne between *t* and *o*, and *t* bearing *r* to *o* will *also* share the feature that *t* will thereby contribute *o* to the truth conditions of all sentences in which *t* partakes. Of course, in virtue of being a *non*-semantic relation, *r* will also have additional features as well (e.g. features pertaining to the production of the term, cognition, the subject's linguistic community, natural environment, etc.). By exhibiting the characteristic feature of reference, the relation *r* in each case will *manifest* reference. Moreover, in ensuring that the job of reference is done, it will thereby realize the fact that *t* refers to *o*.

3.6 Is the account circular?

Viewed from a distance, the pluralist theory may appear to involve an explanatory circle. An object-based correspondence theory, in effect, seeks to explain truth on the basis of reference. But now, according to my pluralist account, the reference relation is characterized by its role which involves truth. So it seems that truth is explained in terms of reference, and reference in terms of truth—a circle!

To unravel this circle, it is important to pay close attention to how the various explanations are supposed to run. Consider how the pluralist would answer the following questionnaire.

Question 1. What explains the truth conditions of a sentence S containing t as a constituent?

Answer 1. The term *t* is assigned an object *o* as its referent, and then the apparatus of semantic composition generates truth conditions on this basis. We can gloss this by saying that *S*'s truth conditions are determined (in part) by *which* object *t* is assigned as referent.

Question 2. Why does (a particular given) term t contribute (a particular given) object o to truth conditions?

Answer 2. At the level of abstraction required for semantic theory (for answering question 1), it is appropriate to cite the fact *that t refers to o.* (For semantic purposes, it is appropriate to appeal to the level of the functional kind.) However, if what's wanted is a metasemantic explanation of this fact, then the answer resides in some non-semantic relation *r* borne between *t* and *o,* which has to do with the production of the term. This relation will have the feature that it can explain why it is *o*, rather than any other object, that *t* contributes to truth conditions.[62]

Question 3. What is the reference relation itself?

Answer 3. It is the relation that essentially performs the function of assigning objects to words in order to generate the truth conditions of sentences.

Question 4. What makes a relation r between t and o count as a realization of reference?

Answer 4. It is simply that *r* performs the role described in answer 2; it explains why *t* contributes *o* to truth conditions.

On closer inspection, we see from this Q&A that truth conditions are *not* explained in terms of reference in a way that's problematic. To be precise, the answer to question 1 does *not* presuppose the answers to questions 3 and 4. To explain the selection of truth conditions, semantic theory does not rest on an answer as to what the *reference relation is* or which *non-semantic relations determine reference.* Rather, all that it requires is an answer as to *what the referents are.* We might gloss this by saying that truth-conditional attributions are explained (at the semantic level) in terms of referen*ts*—not referen*ce*—however it is that they're determined.

Secondly, the determination of the referen*ts* of words (Q&A 2) is explained by their metasemantic grounds: i.e. the particular explanations that appeal to non-semantic relations for each term and referent, taken severally. Again, I claim that these relations are apt to stand on

[62] When we state the question and answer it in a generic form, as we have done here, it will sound bad. "Why is it that *P*?" "Because there is an *r* that has the feature of explaining that *P*." This has the form of the notorious *virtus dormitiva*. However, it only sounds bad because we have attempted to answer the question generically. One of the central claims of the pluralist account is that there is no *general* answer to this kind of question. Rather, in each instance, there will be a particular relation between word and referent that supports this kind of explanation. And when we attend to the specifics, the explanations will not be empty.

their own for the sake of their own explanatory purposes. They are capable of explaining the selection of referents without presupposing a theory of what reference generally is (Q&A 3).[63]

The pluralist account of the nature of reference (Q&A 3&4) enters the stage only *after* all of these other chips have already fallen. It is not designed to contribute to answering the first-order selective questions ('why does *t* refer to what it does?' and 'why does *S* have the truth conditions that it has?'). Instead, it assumes that the answers to these questions have already been given and then it overlays them with an answer to the metaphysical question, 'what is the nature of reference?'.[64] Since, according to this picture, the answer to the nature question presupposes that there are answers to each selection question, *but not the other way around,* there is no overall circle.

3.7 Why be a pluralist?

Near the beginning of this chapter I noted that the traditional approach to the theory of reference has operated under several steep constraints. The traditional theories had the ambition to say, all at once, what reference itself *is*—and in non-semantic terms no less—and what the general conditions are for reference fixation. The theory advanced in this chapter does not share all of these ambitions. Nonetheless, it's high time that we see this as a virtue.

As previously mentioned, the ulterior motive behind the first constraint came from a reductionist program that was fostered by metaphysical naturalism. The reason to analyze reference in *non-semantic* terms, according to this motive, was because it was thought to be necessary for reconciling semantics with a broadly naturalistic worldview.

[63] This feature distinguishes the present view from the interpretationist approach to metasemantics that was developed by Donald Davidson and David Lewis. Much about this will be said in the next chapter.

For now, I note that certain versions of interpretationism threaten to incur an explanatory circle that the functionalist avoids. Both the functionalist and the interpretationist take note of the variety of non-semantic relations (e.g. causation) that terms bear to their referents. Both of them see these non-semantic connections as ingredients for metasemantic explanation. But for an interpretationist such as Sider (2013), claims such as *'t* refers to *o* because *t* bears *r* to *o'* become part of the theory whose overall truth requires explanation by an appeal to the best interpretation. The problem, as pointed out by Simchen (2017), is that *truth* is also part of the overall explananda of the metasemantic claim '*t* refers to *o* because *t* bears *r* to *o*'. Thus it is inappropriate, within one's overall metasemantic story, to appeal to the *truth* of the sentence '*t* refers to *o* because *t* bears *r* to *o*', as one more truth to be interpreted. For the purposes of metasemantic explanation, the first-order claim that *t refers to o because t bears r to o* and its semantic ascent are not explanatory equals. While this may be a problem for Sider's version of interpretationism (or any theory which makes the 'just more theory' maneuver towards first-order metasemantic explanations), it is not a problem for reference functionalism. The reference functionalist explains first-order reference facts—*that t refers to o*—with first-order non-semantic facts—*that t bears r to o*. It does not semantically ascend the explanation.

[64] I am speaking figuratively here. We, as theorists, do not need to *know* what determines the referent of each and every referring term in order to embrace the pluralist theory of reference. Here, as elsewhere, the operant notion of 'explanation' is *metaphysical,* not epistemological.

Well, if naturalism is your concern, then the pluralist theory can reap the reward without the headache of reductionism. That is because a functionalist account of a property or relation is generally considered to be compatible with physicalism provided that all of its realizers are. In this case, the semantic grounding thesis, combined with the functionalist account of reference, guarantees that each instance of reference is realized by some non-semantic relation or another. (Unlike the traditional theories, I make no promise that we can say what they are in each case; but no matter.) So there's a clear sense in which the semantic facts are determined by the non-semantic facts on this view.

But to my mind, there's another advantage to pluralism that is even more important. The theory is *highly versatile* in its approach to first-order semantics and metasemantics. Unlike the traditional theories that rigidly adhere to one-size-fits-all conditions for reference-fixation, the pluralist permits the stories of reference-fixation to operate on a case-by-case basis, and adhere to standards that are appropriate to the domain in question. For this reason, it respects the autonomy of each field of inquiry (cognitive science, philosophy of mind, metaethics, etc.). It allows them to pursue their metasemantic questions according to their own proprietary explanatory aims. And this seems to me to be a sound policy.

It is also difficult to understate how beneficial this is for the metasemantic questions that concern *philosophers.* Philosophical problems abound the moment that we consider reference to abstracta, numbers, moral properties, social kinds, and the self (just to name a few). Given the enormous diversity among these referents, especially considered alongside sundry concrete objects and the theoretical objects of science, a one-size-fits-all theory of reference-fixation is seen to be hopeless. It is therefore desirable for a metaphysics of the *nature* of reference to loosen the leash around the states that *determine* reference. To this end, a pluralist account is exactly what's needed.

With the leash loosened for selective explanation, one might fear that the pluralist is making things too easy and unprincipled. The final worry is that frictionless metaphysics will lead to irresponsible metasemantics. If there are no general principles for reference fixation, then perhaps anything goes.

But on the contrary, there is nothing in the pluralist account that suggests that the remaining tasks for metasemantics are easy. This brings us back to a point made earlier: the pluralist theory allows (indeed, claims) that each selective explanation belongs to a particular explanatory context. (An explanation for my demonstrative '*that*' is part of a broader effort to explain the content of personal-level perceptual demonstratives; an explanation for the numeral '7' is part of a broader effort to explain mathematical reference.) And each context will have its own explanatory aims and constraints. Within a particular explanatory setting, it will be profitable to seek semi-general reference-fixing principles for each term belonging to the kind (however it is that terms are divided up into kinds). Mid-level principles are important precisely because they make our metasemantic explanations robust. For instance, they allow us to extrapolate our explanations to novel terms of a familiar kind, thereby yielding predictions about the primitive semantic facts. This explanatory endeavour is entirely in line with the metaphysics of pluralism.

This now brings us to the final virtue of pluralism: it leaves the division of labour exactly where it should be. By recognizing the legitimacy of a variety of explanatory aims, the pluralist approach allows us to value the mid-level accounts of reference when these accounts succeed on their own terms. Pluralists can thus view the theories of content from cognitive science as partially informing our understanding of reference (e.g. Shea 2018). They can likewise uphold the social constructivist accounts of social kind terms as making another valuable contribution (e.g. Haslanger 2020). There are endless mid-level theories that can be incorporated into the pluralist framework. But according to pluralism, none of these projects needs to take top-down orders from a metaphysician claiming to have uncovered the nature of reference.

Pluralism thus answers the broad, abstract question 'what is reference' (and for that matter, 'what is truth?', when combined with a modern correspondence theory) with a level of generality that is entirely appropriate. It doesn't impose from above; and it incorporates from below all of the many ways that we may seek to explain reference.

Appendix

We are now finally in a position to combine the pieces from chapters one and three into an overall picture of truth. The result is a version of the object-based correspondence theory with a significant element of pluralism. To be specific, the theory takes the truth of sentences (and perhaps propositions) to be grounded in three factors: compositional semantics, subsentential word-world relations, and the way the world is (i.e. objects and their properties). The pluralism of the present chapter targets the second component. It allows for metasemantics—both its means (the factors that explain reference) and its ends (the criteria for explanation)—to vary from domain to domain. Hence, the explanation of, say, mathematical reference can be quite different from the explanation of reference to perceptible objects or social constructs.

From the current literature, the present view appears to be most similar to the theory of Sher (2015 & 2016). Sher also presents a pluralist version of the correspondence theory of truth. Her theory is pluralistic because it allows for the correspondence relation between true statements and the world to vary from domain to domain. However, closer inspection reveals that the *nature* of the pluralism inherent to Sher's view differs from the pluralism inherent to mine.

According to Sher, truth is plural because it need not always consist of an isomorphic correspondence between singular terms and individuals, predicates and properties, quantifiers and sets of objects, and so on (2015: 195; 2016: 200-3). Sometimes it does, and sometimes it doesn't. Presumably, truths about ordinary individuals (e.g. 'Socrates is wise') will consist in the usual semantic relations between singular terms and individuals, and predicates and properties. But for Sher, this is not the case for mathematical truths. Mathematics is different because its worldly subject matter consists of laws between higher-level properties (e.g. the higher-order property of having *n-many instances*), and yet, we report on these laws using singular terms (i.e. numerals—'*n*') and first-order predicates (e.g. 'is odd'). Hence, mathematical truth must consist of an *indirect* kind of correspondence (2015: 200; 2016: 201).

Sher explains this indirect correspondence as consisting of *two* kinds of semantic relations (2016: 201). First, ordinary mathematical language is said to 'simply' refer to *posited* mathematical individuals and first-level properties (e.g. '1' refers to the posited individual *1*; 'odd' refers to the posited property of *being odd*). Secondly, these posited individuals and first-level properties are said to *represent* the higher-order properties that form the ultimate subject matter of mathematics. Mathematical *truth* is then explained in terms of the systematic connections between the posited individuals (and their properties) and the represented higher-level properties (and their laws). So ultimately, what makes a mathematical statement true is its indirect representation of a law between higher-level mathematical properties.

Sher's proposal presents a fascinating solution to a variety of puzzles in the philosophy of mathematics (see 2015: 201-3; 2016: 203-8). Needless to say, it is beyond my scope to assess it here. Rather, my purpose is only to explain how it is distinct from the present view.

As we can see, Sher's theory is pluralistic precisely because it posits a plurality of *semantic* relations. Truth, in her view, is sometimes explained by straightforward reference, and sometimes it's explained by two-step semantic relations involving both simple reference (to posited intermediaries) and indirect representation. By contrast, the theory of the present chapter does not subscribe to a pluralism of semantic relations. On the contrary, it locates the pluralism in the *metasemantic* explanations *of* the reference relation borne between words and their referents. It holds that the reference facts may be determined by a plurality of underlying factors —and that these factors may differ from topic to topic. Since the pluralism of the present view is metasemantic rather than semantic in character, the resulting picture of truth is ultimately distinct from Sher's.

Chapter 4: Reference Functionalism and Metasemantic Interpretationism

4.1 Introduction

Throughout the preceding chapters, my efforts have been to develop a moderately inflationary correspondence theory that relies on a functionalist account of reference. Let's start with a quick recap of the overall picture.

Viewed from one angle, the theory presented here has the same structure as what Donald Davidson derided as the 'building block theory' of truth (1977: 252). It seeks to explain the truth conditions of sentences as the products of the semantic compositional principles and the referential features of words. Moreover, it seeks to ground the referential features of words in certain non-semantic relations between each word and its referent. The theory is thus atomistic in the sense that it breaks down the explanation of the semantic properties for complexes (e.g. sentences) into several piecemeal explanations for the semantic properties of the atoms (e.g. names and predicates). The piecemeal treatment of the atoms makes this picture pluralistic in its understanding of the determination of reference.

However, when viewed from another angle, the theory also offers another sort of unity through its functionalist account of reference. The previous chapter proposed that reference, as a general kind of relation, can be characterized by its role within semantics; specifically, it functions to explain why the truth conditions of sentences depend on the properties of selected objects. The functionalist account also claims that any non-semantic relation can realize reference, provided that it serves to explain why the referent was selected.

The purpose of this chapter is to clarify the nature of this second commitment. The reason clarification is needed is that the functionalist theory appears to come dangerously close to reiterating some of the broad commitments of metasemantic interpretationism. Specifically, it appears to be neighbouring the interpretationist's claim that the facts of reference are fixed by a prior assignment of truth conditions to all of the sentences of the language. In one of his defences of interpretationism, Davidson writes that reference is a "theoretical construct[] whose function is exhausted in stating the truth conditions for sentences" (1977: 255). This comes too close for comfort.

If the functionalist theory collapses into this view, then that would be a disaster. For one, the view presented here claims that truth conditions are determined by the facts of reference. So if the facts of reference are also determined by a prior assignment of truth conditions, then the whole picture results in a circle. We have already addressed this worry in §3.6, but an elaborated response depends on distancing the view from metasemantic interpretationism. Another worry about being associated with interpretationism is that the latter view is notoriously plagued by objections having to do with referential indeterminacy. Given the apparent proximity between the functionalist theory and interpretationism, one might wonder whether the same worries arise for functionalism.

For these reasons, the main item on the agenda for this chapter is to show that the functionalist account is, in fact, quite distinct from interpretationism. To this end, I will sketch the classic interpretationist views of Donald Davidson and David Lewis. The purpose will be to contrast these pictures with the claims of reference functionalism that were outlined in chapter three. In addition, it will be important to discuss various problems of indeterminacy with regard to each version of interpretationism. This allows us to contrast the functionalist's response to the problems of radical indeterminacy.

4.2 Metasemantic Interpretationism

It is not easy to give a universal description of interpretationism, since it is not a specific view, but a large family of approaches to metasemantic explanation. To add to this difficulty, each version has different explananda. Metasemantic questions are generally concerned with the fixation of semantic properties (truth conditions, reference, meaning), but different versions of interpretationism focus on different items of semantic significance. Throughout Davidson's writings, his primary concern is to assign truth conditions to the sentences of a speaker's idiolect, which he regards to be central to the speaker's beliefs (through the attitude of 'holding-true') along with their public assertions (see e.g. 1974). Lewis, on the other hand, produced a few different versions of interpretationism throughout his career. His earlier view (e.g. 1974, 1975) first assigns belief and desire contents (propositions understood as sets of worlds) to a subject, and then subsequently uses these attitudes to interpret the semantic properties of public sentences via conventions. However, in his (1984), he also defends a version of 'global descriptivism', which takes the interpretation of theories to be primary. So overall, there are substantive disagreements over what-gets-interpreted (sentences or attitudes) and what-does-the-interpreting (truth conditions or propositions). We will soon explore these differences in detail.

But before we do, we can still present the interpretationist's approach in a rough schematic form. The most fundamental idea of interpretationism is that semantic properties are somehow conferred by, or grounded in, the canons of interpretation. What makes a semantic item mean what it does *is just that it is interpretable as such* (Simchen 2017: 4).

To put this another way (paraphrasing Williams 2020: 9), it is a truism that:

Item x has semantic content c if and only if x is correctly interpretable as meaning c,

when this is read as a mere biconditional. According to a realist, non-interpretationist reading of this truism, the order of explanation runs from left to right: correct interpretation is grounded in the prior semantic facts. (This echoes the inflationary order of explanation from §2.3.) But for the interpretationist, the order of explanation goes in reverse. Rather than subscribing to a realm of semantic facts that ground the notion of correct interpretation, they see the semantic facts as *arising* out of the facts of correct interpretation.

In light of this metaphysical orientation, there are a few features that follow that are common to the mainstream versions of interpretationism. One that will become especially important for this chapter is the prioritization of sentential semantic facts over subsentential semantic facts. According to each interpretationist theory, the interpretation of sentences is fixed before the interpretation of subsentential expressions. This means *inter alia* that, in their view, the facts of reference are determined in part by a prior assignment of truth conditions. We might put this by saying that reference is partly grounded in truth. One reason for prioritizing sentences was provided by Donald Davidson and ultimately stems from Quine (1960). It is that, within the context of interpretation, the interpreter first has empirical access to the sentences 'held true' by the subject, before they are in a position to parse any subsentential structure (Davidson 1977: 252). So the canons of interpretation dictate that sentences must come first. (In the final section, I'll comment on the possibility of the interpretationist dropping this claim.)

Another feature that is common to all interpretationist theories is their *holistic* view of semantic determination. When interpreting a subject's speech and belief system, we must take into account all of the regularities in their behaviour (both verbal and non-verbal) to find the best overall interpretation. (This follows form another Quinean doctrine: confirmational holism.) Combining this fact with their claims about semantic determination, we get the result that, for the interpretationist, semantic facts are determined holistically. This means that the unit of semantic interpretation must be the subject's entire language or belief system. The semantic properties of individual words cannot be determined in isolation.

We'll see how these common features are implemented in each version of interpretationism. To add more substance to the view beyond these vague, general remarks, we need to attend to the details of each interpretationist theory.

4.2.1 Davidsonian interpretationism

Let's start with Davidson, since he is the one who is largely responsible for the framing of the interpretationist project.

The framing that I'm referring to is the famous thought experiment concerning radical interpretation. The idea is to imagine ourselves as encountering a language community for the very first time. We suppose that this community speaks a language that is entirely unfamiliar to us. We may even stipulate, for the sake of illustration, that their language has no phylogenetic relation to our own (maybe its speakers are martians). Now imagine that we come across an individual from this community; let's call him Paul. And imagine that we follow Paul around and observe his behaviour. We observe Paul interacting with his environment and his fellows while making various utterances that we don't understand. Like a detective, we take meticulous notes and record all of these interactions. Once we've collected enough data, our task is to decipher what Paul *means* by the words in his idiolect. Let's call his language *L*. We must go from *observing his overt behaviour* to *understanding L.* Much of Davidson's work throughout the seventies was chiefly devoted to this puzzle.

For Davidson and the other interpretationists that followed, this question of how we can decipher a language from the position of a radical interpreter takes on a particular metaphysical significance. This isn't just a question of local epistemology, as if the aim is to uncover the previously determined semantic facts. According to Davidson, "what a fully informed interpreter could learn about what a speaker means is all there is to learn; the same goes for what the speaker believes" (1983: 148). The idea of radical interpretation is thus a dramatic way to reveal the metaphysical *determinants* of the semantic facts.[65]

Davidson's solution to the problem of radical interpretation can be divided up into two parts. First, we need to specify the appropriate *structure* of a semantic theory for a speaker's (Paul's) language *L*—that is, we need to say what the final product should look like. And secondly, we need to outline the appropriate procedure for determining *which* semantic theory is correct—that is, we need to specify what the evidence is and how it works. Bear in mind that on Davidson's outlook, answering this latter question will also serve the dual purpose of giving a metaphysical foundation for semantics.

With respect to the first question, Davidson proposes several constraints on what an appropriate semantic theory should look like. The first one sets the stage for his eventual theory. According to Davidson, the task of semantic theory should *not* be construed as a matter of assigning *entities* to each expression of the object language to serve as 'the *meanings*' of the expressions. In his view, semantics is not concerned with postulating propositions, or pairing expressions with objects; it's not a branch of specialized ontology. That is because, according to

[65] "What is it for words to mean what they do? ... I explore the idea that we would have an answer to this question if we knew how to construct a theory satisfying two demands: it would provide an interpretation of all utterances, actual and potential, of a speaker or group of speakers; and it would be verifiable without knowledge of the detailed propositional attitudes of the speaker." (Davidson 1984: introduction)

Davidson, the reification of meanings into entities isn't necessary for the central task of semantics, which, for him, is interpreting one another's speech (1967: 307). Instead, all that we need in order to understand another person's speech is to possess a sort of 'interpretation manual' that could be used to interpret them. In order *for us* to understand (say) Paul's speech, what we need is to formulate *in our language* (a metalanguage, *ML*) a theory that allows *us* to interpret each sentence of *Paul's language* (the object language *L*).

The next step for Davidson is to state what kind of theory, formulated in our language *ML*, would constitute a semantic theory for an object language *L*. Davidson suggests that such a theory should be finitely axiomatizable and compositional since that would allow us to understand a potentially infinite number of the speaker's utterances. He also suggests that the theory should be extensional, in order to dodge the difficulties that arise with intensional contexts (1967: 309). To meet all of these demands, Davidson famously (and notoriously) proposes that the answer lies in a Tarskian definition of truth for *L* (310). The idea here is that if we could construct a comprehensive theory that could tell us when the subject's utterances are *true*, then that would be tantamount to understanding the subject's speech. (This is a *very* bold claim. It is also one of the main sources of controversy within the Davidsonian picture. See Soames (2012) for a sustained critique. But it will not be my purpose to engage with this controversy here.) Our semantic theory thus needs to deliver, for each sentence 'S' of the subject's language *L*, a theorem of the form *'S' is true iff P*. Once again these are called the T-sentences. For if we could generate all of the T-sentences for Paul's language, then by Davidson's lights, we will have uncovered all of the facts of meaning.

It is crucial for Davidson's picture that our definition of truth for *L* takes the form of a Tarskian theory because that is the key to meeting the finitary and compositionality constraints. The Tarskian theory provides the means to produce all of our desired T-sentences from only a finite stock of axioms. Basically, if we can find in the language *L* the same structure as the language of first-order logic, then we can give a simple definition of *truth-in-L,* as follows. First, we specify the axioms that give reference clauses for *L*'s singular terms and satisfaction clauses for *L*'s predicates. These axioms will have the forms *'a' refers to b* and *x satisfies 'F' iff x is H.* Then we provide recursive rules for generating the truth conditions of complex sentences from the semantic contributions of their parts. The end result will be a complete recursive definition of *truth-in-L.*

This tells us what a semantic theory for Paul's language should look like. In short, we're looking for a theory (expressible in our metalanguage *ML*) that specifies the reference and satisfaction clauses for Paul's singular terms and predicates, so that we can derive the truth conditions for all of Paul's potential utterances *à la Tarski*. Following this, the second component of the Davidsonian picture concerns the *evidence* and *procedure* for determining which semantic theory is correct.

Throughout his work, Davidson stresses that this is a thoroughly *empirical* question. Ultimately the evidence for our semantic theory ought to consist in Paul's verbal behaviour, which should be specifiable without presupposing any meanings. All we can do (at first) is collect data about the noises that Paul makes in the various contexts that we find him.

In order to convert these observations into evidence for a semantic theory, the first step we should take, according to Davidson, is to identify which sentences the speaker holds to be true. It is Davidson's belief that we can observe speakers bearing the attitude of *holding the sentence 'S' as true* before we can tell what 'S' means. If Davidson is right about this, then we can let the occasions that prompt this attitude be our fundamental data for semantic theory. The basic idea is to observe whenever Paul holds a sentence 'S' to be true, and observe which circumstances coincide. Our hope is that the accumulation of such observations will provide a clue as to what Paul means by 'S'.

Suppose, for instance, that we regularly observe Paul hold the sentence 'S' as true whenever it is raining, and we never observe him take this attitude when it isn't raining. Given enough of these observations, we might conjecture the following T-sentence: *'S' is true-in-L if and only if it is raining*. We might even say that this T-sentence is *well-confirmed* by the available empirical evidence. If so, then we would want our definition of truth for *L* to entail this T-sentence as a theorem. We would then have some empirically grounded *semantic* evidence to calibrate our overall semantic theory. Continuing this line of thought, we should try to gather as many of these empirically-confirmed conjectured T-sentences as we can. For the more empirically confirmed T-sentences we have at our disposal, the more semantic data points we have to construct our semantic theory.

An obvious problem arises if we were to naively assume that these empirically-confirmed T-sentences have evidential value. Namely, that the states of the world only provide evidence for what a speaker means *if* we assume that the speaker has largely true beliefs. If the speaker has wildly false beliefs, then our conjectured T-sentences are unlikely to capture what they mean by their utterances. (Imagine that Paul has a conspiratorial belief about ghosts sprinkling water over his head. In that case, his 'S' utterances may coincide with rain, but this coincidence will not be a reliable guide to what he means.) Our problem is then compounded by the fact that we have no way of learning what a speaker believes if we do not know what their words mean. But we cannot know what their words mean unless we have some knowledge of what they believe. So in order to uncover what Paul means by his speech, we are going to have to interpret his beliefs and meanings *simultaneously.* But how do we break into this circle?

Davidson's solution is to invoke his famous 'principle of charity'. This principle licenses us to assume *a priori* that the speaker has largely true beliefs—or to be more precise, that they have beliefs that are largely *true by the lights of the interpreter's worldview.* As Davidson presents it, the principle has three components.[66] First, as the interpreter of the speaker's beliefs and meanings, we should seek to attribute logically consistent beliefs to the speaker as much as possible (although we can allow exceptions if we need to). Secondly, we should also seek to credit the speaker with as many true beliefs (true by our lights) as possible; that is, we should, as often as possible, pair *held-true* sentences on the left-hand side of our conjectured T-sentences with what we believe to be true sentences on the right-hand side. Davidson writes that this procedure is justified by the fact that "disagreement and agreement alike are intelligible only against a background of massive agreement" (1973: 324). And thirdly, we should prioritize the attribution of true beliefs when the beliefs in question are about the speaker's immediate environment. (It's safe to assume that Paul is more likely to be accurate about the weather than about fundamental physics.)

With the principle of charity in hand, we now have some license to infer that *'S' is true-in-L if and only if it is raining,* given that Paul tends to hold 'S' as true whenever it's raining. The principle thus gives us the foothold we need to amass our fundamental semantic data in the form of empirically-confirmed conjectured T-sentences. Once we have assembled enough of these T-sentences, we are then in a position to start interpreting the subsentential expressions of *L.* In order to do this, we must first detect the logical vocabulary of *L* and parse its syntax. After we have that sorted out, we will then be in a position to interpret the singular terms and predicates. We do this by hypothesizing R-sentences and satisfaction clauses for the subsentential parts of Paul's language. These hypotheses will take the forms *'a' refers to b* and *x satisfies 'F' iff x is H,* and they will constitute the base clauses for our Tarskian recursive definition. Given these, we can generate a T-sentence for each sentence of Paul's language.

From then on, our method for interpreting Paul is essentially that of hypothetico-deductive confirmation. Our hypothesized R-sentences and satisfaction clauses will generate testable predictions in the form of T-sentences. We can then check to see whether Paul's speech behaviour confirms or disconfirms our predictions; we observe whether Paul holds various sentences to be true in the circumstances that we predict. But of course, like any other empirical methodology, we can never fully confirm or disconfirm a semantic theory by any single observation. We must operate holistically. We aim to construct the theory that has the overall best fit to our linguistic data while abiding by the principle of charity (attributing true belief and logical consistency as much as possible) and acknowledging that we will need to make trade-offs to achieve the best overall fit.

Since our overarching concern is the metaphysics of semantic facts, it is worth dwelling on the role that the semantic facts play within the Davidsonian framework. The crucial point is that for Davidson, theories of meaning are instruments, which we construct for the purpose of

[66] The presentation of the principle evolved throughout Davidson's career. The version that I'm outlining, which gives priority to perceptual beliefs, is found in his (1983).

interpreting one another to facilitate interpersonal communication. Bearing this in mind, consider the location of the 'facts' of reference and satisfaction within the Davidsonian scheme. For him, the fact (if we can even call it a fact) that Paul uses '*a*' to refer to *b* is not to be understood as reporting any objective relation between Paul's cognitive situation and the object *b* (Davidson 1977). Reference is not explained directly by non-semantic (e.g. causal) relations between Paul and the object *b*, and his theory does not assign 'any empirical content directly to relations between names or predicates and objects' (1977: 255). Rather, for Davidson, each fact of reference is an artifact of the Tarskian truth definition that *we,* the interpreters, have constructed to understand Paul's speech. It is *because we* interpret Paul's use of '*a*' as referring to *b*, as a clause within an interpretational theory that offers the best overall explanation of Paul's behaviour and rationality, that determines this to be a fact. This means that, in Davidson's view, the facts of reference and satisfaction are determined holistically, by the total evidence afforded by all of Paul's speech behaviour. It also means that these facts depend on the perspective of the interpreter. Without *our* interpretation of Paul, there would be no further stance-independent facts about the semantic properties of his speech. The facts of reference and satisfaction are essentially *perspectival.*

4.2.2 Lewisian interpretationism

The other famous variety of interpretationism was developed by David Lewis. In fact, Lewis delivered two substantially different versions. One is found in his (1969), (1974), and (1975) and was concerned to address roughly the same problem as Davidson: outlining the canons of interpretation in the context of radical interpretation as a way of explicating the grounds for semantic facts. The other version is his 'global descriptivism' with the eligibility constraint, which was outlined in his (1984) and developed in response to a radical indeterminacy problem advanced by Putnam. I will discuss each one of them separately.

For our purposes, the best way to outline Lewis' first version of interpretationism is to contrast it with Davidson's. (Lewis explicitly makes this comparison in his 1974.) In brief, when the radical interpreter follows Davidson's guidelines, they proceed in this order. First, they uncover which (not-yet-interpreted) sentences Paul holds to be true. Then they use this evidence to interpret Paul's beliefs and sentential meanings simultaneously. Finally, they use the assignment of sentential meanings to discern an assignment of subsentential meanings.

Lewis has two objections to this order. For one, he says that it places far too much emphasis on the subject's *verbal* behaviour as a basis for interpreting their beliefs. But verbal behaviour is only a small part of overall behaviour, and it's better to take heed of the rest when interpreting a subject's propositional attitudes (1974: 341). In addition, Lewis says that

Davidson's method under-utilizes the role of linguistic conventions in determining the meanings of the subject's speech in a public language (1974: 341). For these reasons, a different approach is called for.

What is distinctive of Lewis' brand of interpretationism is that, unlike Davidson, he gives priority to the interpretation of beliefs, desires, and the rest of the mental attitudes before public language. He calls this the 'head-first' approach. Hence, in the context of interpreting a particular individual, Paul, our first step in the Lewisian approach is to assign to him a set of propositional attitudes. Our evidence for these attitudes is no longer his uninterpreted utterances. Instead, we look at the history of evidence that has been given to him as well as his overall behaviour "given in physical terms" (1974: 337). This is supposed to be sufficient to assign a set of propositions and attitudes to our subject by using the principles of *charity* and *rationalization* (336–7). Roughly speaking, we endeavour, as much as possible, to optimize the rationality of the assignment of Paul's beliefs and desires to his evidence and behaviour, according to what is reasonable and valuable. At first pass, the attitudes that we assign may be regarded as relations to propositional contents, conceived as sets of possible worlds.[67] (Unlike Davidson, Lewis is not bothered about 'reifying' meanings (Lewis 1975: 690). If we need to, we can elaborate the picture by including centred worlds for *de se* attitudes and degreed beliefs and desires, but let's ignore these complications for a simplified presentation.

Once the beliefs and desires of a subject are determined and settled, we can then make full use of them to interpret the meanings of the subject's public sentences by appealing to the entire linguistic community that the subject is party to. This is the project that Lewis undertakes in his book *Convention* (1969) and his (1975). Each of these works contain an enormous amount of detail, and so we can only give them the most cursory of glosses here. In its most basic form, Lewis' idea is that the public meanings of the utterances of a population are determined by certain conventions. In his (1975), he specifies those conventions as *truthfulness* and *trust*.[68] Let's say that a *language L* is an assignment of meanings (propositions) to a system of sentences. The fact that one language *L* is used by a population *P*, as opposed to any other language, is determined (according to Lewis) by the twofold facts that:

- there prevails in *P* a convention of aiming never to utter a sentence of *L* unless it's believed to be true in *L*.
- there prevails in *P* a convention of imputing truthfulness in *L* to others, and thus to tend to respond to another's utterance of any sentence of *L* by coming to believe that the

[67] This is the line that Lewis takes in his (1969) and (1975). In his (1974) he allows us to interpret the attitudes by an assignment of non-reified truth conditions, as a capitulation to Quine and Davidson. But his 'official' view is the propositionalist one.

[68] In his earlier works (1969, 1974), Lewis only discuss the convention of truthfulness. He explains in his (1975) that *trust* should be included as well.

uttered sentence is true in *L* (1975: 684).

For example, that Paul's utterance *u* means *it's raining* will be determined, on the Lewisian picture, by the general conventions prevailing within Paul's community of uttering *u* only if one believes it to be raining, and of forming the belief that it's raining whenever others make the same utterance. Finally, Lewis analyzes conventions as *certain regularities in action* that arise in response to coordination problems. For our purposes, suffice to say that a convention occurs whenever a population regularly performs a type of action, they each desire and believe that everyone conforms to this regularity, but they could have arbitrarily chosen some other action to achieve roughly the same end. (Each English speaker associates 'it's raining' with the proposition *that it's raining*, we desire and believe all other English speakers to do the same, but we could have arbitrarily chosen some other sentence to pair with this proposition.)

Not unlike Davidson, the last phase of the Lewisian interpretation procedure is to parse the subsentential meanings of the subject's language. In Lewis' terminology, this is called assigning a 'grammar' to a language. This consists of parsing the syntactic structure, the rules of semantic composition, and the assignment of denotations (1975: 690). However, unlike Davidson, Lewis takes a more liberal approach as to how a grammar may be shaped. Whereas Davidson rigidly adhered to the Tarskian structure, Lewis allows for different parsings of syntax that are more faithful to natural language syntax (ibid). He also prefers an assignment of Carnapian intensions to basic subsentential expressions (names are assigned functions from worlds to individuals; predicates are assigned functions from worlds to classes of individuals, and so on).

The rules for assigning a grammar to a language, according to Lewis, are (once again) a matter of best overall fit. Having already determined the contents of whole sentences, understood as coarse-grained sets of possible worlds, the primary constraint on the choice of grammar is that it generates the right contents (1974: 339, 342). We thus fill out the meanings of each word with the aim of reconstructing the already-determined meanings, including truth conditions, of sentences. So again, much like Davidson, Lewis places truth before reference (*inter alia*) in the order of explanation.

4.3 Functionalism vs interpretationism: the grounds and nature of reference

This gives us just enough to see the superficial similarity between the interpretationist and functionalist points of view. Like the functionalist, the interpretationist characterizes reference by its role within semantic theory. For Davidson, reference is a matter of assigning semantic values to words for the purpose of interpreting a subject's overall language using a Tarskian truth definition. The reference clauses serve as the basis for a recursive, systematic theory of meaning. Moreover, in Lewis' view, reference is also a matter of generating truth conditions within some appropriate formal semantic theory. Either way, reference is characterized by what it does to the truth conditions of sentences. And on this point, the functionalist agrees.

But the agreement does not extend very far. There are crucial differences as to how the functionalist understands both the determination and function of reference. Recall, from the previous chapter, that the functionalist provides distinct answers to the questions about the nature of reference and the grounding or selection of particular reference facts. To contrast their view with interpretationism, we ought to consider both of these questions separately.

Let's start with the selective explanations of reference. In order to approach this topic, it is best to focus on one or two concrete examples of reference relationships. To this end, the examples from chapter two will serve just fine.

Case 1 (Demonstrative) One very foggy night, my partner and I take a long walk on the beach. In the distance, there's a dim light, mostly obscured by the fog. My partner looks at it and points towards it and says '*that* is a lighthouse'. The referent in question is a certain object *o*, which is the cause of the light.

Case 2 (Proper name) Continuing on our walk, our conversation turns to natural wonders. My partner says in a matter-of-fact way, 'Kilimanjaro is the world's tallest free-standing mountain'. Of course, by her use of 'Kilimanjaro' she is referring to Kilimanjaro.

The aim of selective explanation is to explain why the referent was selected to the exclusion of other objects. Let's say that in these cases, we are concerned to explain (i) why *my partner's use of 'that' refers to o rather than the distant mountains,* and (ii) why *my partner's use of 'Kilimanjaro' refers to Kilimanjaro rather than Everest.*

As outlined in the previous chapters, the metasemantic explanations provided by the functionalist will fall under the heading of productivism. To reiterate, this means that the determinants of reference are found within the circumstances that lead up to the production of each referring term. The fact that my partner's use of '*that*' refers to *o* (rather than the distant mountains) is grounded in the intentional and perceptual relationships that she bore to *o* as she uttered the token. The fact that her 'Kilimanjaro' refers to Kilimanjaro (rather than Everest) is explained by a chain of social deference that traces back to Kilimanjaro rather than Everest. These relations are cognitive and social; they involve the intentional states of my partner and others. However, these other intentional states are also (presumably) susceptible to selective explanation. Their referential features should ultimately be grounded in non-semantic relations, such as causation.

There are a couple of points that must be highlighted. First, notice that the productivist orientation is geared towards providing selective explanations for each term and referent *taken on their own*. On this view, the basic target of metasemantic explanation is particular reference relationships, such as my partner's use of '*that*' to refer to *o* and her use of 'Kilimanjaro' to refer to Kilimanjaro. The explanations offered by the functionalist may thus pertain to the specific circumstances of each referring term. In this sense, the functionalist's metasemantics is piecemeal. We will see shortly how this contrasts with interpretationism.

The other thing to observe is that the functionalist, *qua* productivist, does not mention the truth values of a term's host sentences within the selective explanation of reference for that term. To explain why my partner's token of 'Kilimanjaro' refers to Kilimanjaro, we only need to look at the term's productive history (e.g. her referential intentions, the chain of historical uses within her community). We do not mention whether my partner used it to say something true or false. This is by design since the functionalist account is meant to be compatible with the inflationary order of explanation (chapter two) and the object-based correspondence theory (chapter one). Truth conditions are thus to be determined by the referential contributions of words, and so the selection of referents cannot be explained in terms of truth.

Both of these features contrast with the interpretationist's approach to selective explanations of reference. For one thing, we have seen both Davidson and Lewis hold that truth precedes reference in the order of metasemantic determination. For them, a reference fact is determined by its potential to contribute to the best overall theory that captures the predetermined meanings of sentences, including their truth conditions. In order for this to work, the truth conditions of sentences must be settled prior to the facts of reference. This stands in stark contrast to the functionalist's favoured orientation.

The second contrast stems from an interpretationist doctrine that I will call *metasemantic holism*. For the interpretationist, the primary target of metasemantic explanation is not the individual reference facts *per se*, but rather, it is *reference schemes for an entire object language.* A reference scheme for an object language *L* is an assignment of extensions (or intensions) to all of the referring terms of *L*. Thus, for the interpretationist, the fundamental metasemantic question takes the form:

(MQ) What determines that a reference scheme *R* is true of the object language *L*?

Moreover, the answers to this question must be features that are attributable to *overall reference schemes.* The interpretationist may appeal to such global virtues as *overall simplicity, overall explanatory depth, maximization of the appropriate kinds for kind terms, etc.* Since these features target reference schemes *as a whole,* the whole is explanatorily prior to its parts.

This metasemantic holism shapes the interpretationist's approach to individual selective explanations of reference. Consider the fact that my partner's use of 'Kilimanjaro' refers to *Kilimanjaro* and not *Everest.* To explain this fact, the interpretationist cannot attend solely to the cognitive, social or causal relations that pertain directly between my partner's token and its referent. Rather, they examine the entirety of two competing reference schemes, R_1 and R_2. Let R_1 be a reference scheme for my partner's language that entails, *inter alia,* that Kilimanjaro is assigned to 'Kilimanjaro', and let R_2 be a reference scheme that entails, *inter alia,* that Everest is assigned to 'Kilimanjaro'. The explanation for why her use of 'Kilimanjaro' refers to Kilimanjaro rather than Everest, for the interpretationist, is that R_1 beats out R_2 on the score for being the best overall interpretation. In this way, reference schemes and their global virtues take priority over particular reference facts viewed in isolation.

This holistic approach to metasemantics follows naturally from the interpretationist's outlook. According to interpretationism, the facts of meaning arise out of the circumstances that determine the correct interpretation of another's language. But the evidence provided by these circumstances for a given interpretation will only confirm it holistically. In the context of radical interpretation, the interpreter does not confirm or disconfirm an isolated semantic hypothesis on its own. It is only when taken as part of a larger interpretational scheme that a semantic hypothesis affords confirmation. Likewise, according to metasemantic holism, the determinants of a particular reference fact do not pertain to the reference fact in isolation; rather, each reference fact is determined by its participation in an overall assignment of referents that's determined to be correct by its global properties.

For the reasons presented above, we can conclude that the functionalist, *qua* productivist, has a sharply different stance from the interpretationist regarding selective explanations of reference. For that kind of explanation, there should be no temptation to mistake the two views. But perhaps the functionalist will sound more like an interpretationist when they propound their claims about the nature of reference. After all, both the interpretationist and the functionalist characterize reference by its role within semantics. So on this point, there may appear to be more overlap.

However, closer inspection reveals that the two have very different approaches for characterizing reference by its role.

Most notably, the functionalist assigns reference a fairly direct role in explaining *truth.* For the functionalist, reference is the relation that assigns objects to words *for the purpose of determining truth conditions.* This role is explanatory. Reference has a part in *explaining* the truth values of sentences by rendering them dependent upon the properties of the referents of subsentential words. This explanatory role dovetails with the object-based correspondence theory of chapter one.

To be fully explicit, according to functionalism, reference is characterized by what it does in the overall project of explaining truth based on how the mind and language connect with the world through the simple components of thought and speech (singular terms, concepts, etc.). We see this role exemplified in such mundane explanations as *'that is a lighthouse' is true because 'that' refers to o and o is a lighthouse.* But we also see it exemplified in other, less platitudinous accounts. For example, we see it in the cognitive scientist's effort to explain successful and unsuccessful actions by invoking true and false intentional states, which in turn are explained by the symbol-object relations that determine sub-propositional reference. The point is, not all explanations that invoke reference are platitudinous. It's enough that there are some broad explanatory environments where reference has a role to play. And in those environments, reference earns its keep not by being reducible to non-semantic relations but by facilitating genuine explanations.

The interpretationist, by contrast, does not assign reference a direct role in explaining truth. Indeed, for both Davidson and Lewis, *truth* is the more central explanatory notion. Reference is less fundamental. According to Davidson's version of interpretationism, the chief purpose of reference is that it serves as part of an overall explanation of the compositionality of meaning (Davidson 1967). For Davidson, reference (for a language) is an artifact of a Tarskian truth definition (for that language), the overall role of which is to provide a finite means for capturing the infinitely many facts of meaning. The situation is similar for Lewis, except he doesn't require the explanation of semantic composition to take the specific Tarskian form. For each of them, truth is invoked within certain crucial steps of the interpretation process to illuminate meaning. But truth itself is not the proximal explanandum of this process. At best, truth is only indirectly illuminated once its connections to meaning, belief, and behaviour are displayed.

It follows that, for the interpretationist, the dependence of truth on reference is, at best, highly indirect. The best overall interpretation of a language will take the form of a compositional theory that assigns referents to words and truth conditions to sentences. And since it must be confirmed holistically, its reference assignment will have some downstream effect on determining it to be the best (*a fortiori,* determining truth conditions). But this is all very different from how functionalism envisions the connection between reference and truth. For the functionalist, truth is an immediate explanandum of reference.

So, to summarize, the functionalist and interpretationist ascribe opposite priorities to truth and reference. The functionalist takes reference to be the more fundamental, immediate explainer of truth, and the interpretationist takes truth to be the more fundamental explainer of meaning, through an apparatus that involves reference. Since they both characterize reference by the roles they respectively ascribe it, it follows that their accounts are distinct.

To further press this point, it may be helpful to draw an analogy to the divide between psycho- and analytic functionalism in the philosophy of mind. According to psycho-functionalism, special science notions like *pain* (for example) are sanctioned by their explanatory potential for special scientific inquiry. They need not be reducible to lower-level notions to be legitimate (Fodor 1974). As for what *pain itself* is, the psycho-functionalist will appeal to *what pain does* within the cognitive economy of pain-feeling organisms (Fodor 1968). These functions are to be discovered empirically by cognitive science. We need not know, before empirical research, the precise nature of the functions that are essential to pain.

This contrasts with analytic functionalism, which claims that *pain* (and other mental states) is already definable, without the need for further empirical inquiry. To define pain, we consult the *folk psychology* theory *T* which describes the causal interconnections between pain, other mental states, sensory inputs, and behaviour. With this theory in hand, we can define pain descriptively (using the Ramsey-Lewis method) as *the state that occupies the pain role within theory T* (Lewis 1970, 1972). This is taken to be *a priori* since the theory *T* is supposed to implicitly define the folk concept of pain. Once pain is so defined, it will then be an empirical matter as to which neural states satisfy the description (and thereby 'realize' pain). But *pain as such* is analyzable through the folk theory *T*. Notice that, unlike psycho-functionalism, the analytic functionalist's characterization of pain is ineliminably theory-first. The theory *T* is indispensable to the explicit definition of pain.

I want to suggest that the functionalist theory of reference, as I envision it, has a similar relation to semantics and metasemantics as the psycho-functionalist does to cognitive science. To be specific, reference functionalists need not determine, in advance, what a compositional theory of truth conditions must look like. Nor need they say, in top-down fashion, how exactly reference is determined for each kind of term. Instead, they take reference as an unreduced explanatory notion that's legitimized by its potential to explain how truth conditions are determined. As for what reference itself is, the reference functionalist will appeal to *what it does* for systems of representation—thought and language. The precise character of its roles may be a matter of ongoing inquiry.[69]

Whereas reference functionalism stands to psycho-functionalism in this analogy, interpretationism stands to analytic functionalism. As noted above, analytic functionalism does not take pain to be an irreducible property that earns its place in our theorizing by its utility for developing empirical theories. On the contrary, it takes the theory surrounding mental properties as *given*, and uses that theory to construct explicit definitions. Like analytic functionalism, the interpretationist also gives semantic theory a certain priority over reference. For them, the semantic interpretation of a language *L* that assigns truth values to sentences, along with the

[69] Note the hint of *realism* that's inherent to this view. Reference may have roles and constraints that extend beyond the platitudes of folk semantics.

formal apparatus for representing semantic composition, constitute the 'theory' that's given prior to reference. Given such a theory, they can then define reference explicitly in the same style as the analytic functionalist: *reference-for-L is the relation that occupies the reference role within the semantic interpretation of L.* And given such a definition, they may even find 'realizers' within the non-semantic relations that satisfy this description. But since the overall theory comes first (the truth value assignment and composition principles), these realizers would play a marginal role in explaining truth. (Again, since interpretations are confirmed and determined holistically, we can't say that reference and its realizers play *no* role in determining truth conditions. But the role is highly indirect.)

This analogy with psycho- and analytic functionalism is illuminating in several respects. Most of all, it underscores how reference functionalism and interpretationism involve different methods for characterizing reference by its role within semantics. But in addition to this, it also points in the direction of a kind of problem that notoriously plagues both analytic functionalists and interpretationists alike. Owing to its theory-first methodology, analytic functionalism takes our folk theory as the final word for determining the extension of our concept of *pain.* But a worry immediately arises: what if our folk psychology isn't precise enough to carve out a determinate set of realizers for the role of pain? In that case, it would leave the extension of pain underdetermined. Now, as we'll see in the following sections, the interpretationist faces a similar problem.

4.4 Truth-to-reference indeterminacy

The foregoing distinctions are important not only because they clarify the theoretical space, but also because they overlap with the problem of radical semantic indeterminacy.

The threat of semantic indeterminacy arises whenever we are considering the metasemantic underpinnings of some fact of reference. The worry for any theory is that it cannot deliver a decisive referent for a term where intuitively we would think that the matter is settled. Some degree of semantic indeterminacy is inevitable and benign. For example, there may be no fact as to whether 'Kilimanjaro' refers to a certain mountain including a pebble that's soon to be lodged in a hiker's shoe, or whether it refers to the same mountain minus the pebble. But some kinds of semantic indeterminacy are unpalatable. Take, for example, the fact that my partner's use of '*that*' refers to *o* rather than the distant mountains. A credible metasemantic theory ought to deliver the result that this is really a fact. If a theory leaves the matter unsettled, then that's a serious mark against it.

We have already seen how the functionalist approaches this problem from the bottom up. For the selection of referents generally, the view takes them to be explained ultimately by the non-semantic relations between the speaker and their referents, and these explanations may be specific to the term and referent in question. For a demonstrative like my partner's token '*that*', we rehearse a familiar story about the speaker's intentions and perceptual states, which ultimately get cashed out in terms of real cognitive and causal relations borne to the referent. Unlike the interpretationist, the functionalist takes this fact of reference to be entirely settled by the factors that lead up to the term's production. In particular, it is settled *prior* to the semantic properties of the sentences that it partakes in, such as '*that's* a lighthouse'. This follows from the commitment to the inflationist order of explanation from §2.3.

Whereas the functionalist, *qua* productivist, seeks to quell the threat of radical semantic indeterminacy from the bottom up, the interpretationist's top-down approach begets its own unique problems. Notoriously, radical semantic indeterminacy presents a formidable challenge to the interpretationist's view. For this reason, it is important to distinguish functionalism from interpretationism to distance the functionalist from these objections.

Near the end of his (1975), Lewis identifies three points at which one might worry that his scheme would fail to deliver a determinate set of semantic facts. First, a subject's behaviour could fail to determine a unique set of propositional attitudes, even when the principles of charity and rationalization are utilized. Secondly, the community's linguistic conventions might fail to determine an assignment of contents to sentences. Finally, the whole range of sentence contents might fail to determine a unique assignment of meanings to the subsentential expressions. In this section, we investigate the third kind. This is a worry that is (roughly) shared by Davidson's view, whereas the functionalist is immune.

When it comes to the third kind of indeterminacy, Lewis is surprisingly unfazed by it. In his (1975), he writes,

> Unfortunately I know of no promising way to make objective sense of the assertion that a grammar is used by a population *P*… I do not propose to discard the notion of the meaning in *P* of a constituent phrase… To propose that would be absurd. But I hold that these notions depend on our methods of evaluating grammars, and therefore are no clearer and no more objective than our notion of a best grammar for a given language. (1975: 691)

Davidson takes a more radical line. Concerning the assignment of meanings to subsentential expressions, he says,

> We don't need the concept of reference; neither do we need reference itself, whatever that may be. For if there is one way of assigning entities to expressions (a way of characterizing 'satisfaction') that yields acceptable results with respect to the truth

> conditions of sentences, there will be endless other ways that do so as well. There is no reason, then, to call any one of these semantical relations ‘reference’ or ‘satisfaction’. (1977: 256)

It is important for us to see how both of their positions lead to these conclusions.

The principal suspect for a source of these consequences is their mutual commitment to a truth-first order of metasemantic determination. Within Davidson’s interpretation procedure, the interpreter first assigns truth conditions to the speaker’s sentences before they begin to parse any subsentential meanings. Davidson claims that this is the price we pay for making our semantic theory empirical; for it is *sentences*, not words, that are the locus of empirical confirmation (1977: 252). On a similar token, Lewis’ system assigns contents (coarse-grained propositions) to sentences of a language before it assigns a ‘grammar’ to the lexicon. For Lewis, this is because the prevailing linguistic conventions that determine meaning pertain, in the first instance, to the sentences of a language.

Insofar as the interpretationist is committed to having sentential meanings determined prior to subsentential meaning, they open themselves to the following objection. (Later we will discuss the options for the interpretationist to drop this commitment.) Perhaps it is possible to keep the entire range of sentence meanings fixed (for a language) while varying the semantic values of the singular terms and predicates (for that language); if so, then sentential meanings will not, by themselves, determine the values of the subsentential expressions. Let’s call this ‘truth-to-reference indeterminacy’.

One way to make this worry vivid is the model-theoretic argument from Putnam (1981).[70] The traditional target of this argument is *global descriptivism*—the view that meanings are determined by maximizing the truth of a global theory. But the argument can also be brought to pose an obstacle for any view that claims that truth is fixed prior to reference. Here is the argument in its most basic form.

Suppose that our speaker, Paul, has issued a set of sentences in his language which he takes to be true. Let’s further say that we have parsed the standard first-order logical vocabulary of his languages (the connective and the quantifiers) and we can represent the syntax of his sentences in first-order form. We thereby determine that his ‘theory’—the set of sentences that he holds true—is syntactically consistent. Since we have already interpreted the logical vocabulary, it remains for us to interpret the lexicon (the names and the predicates).

Given some standard results from model theory, the fact that Paul’s theory is consistent implies that it has a model. That is, there must be a set of things such that an interpretation of the names and predicates will render each sentence of Paul’s theory true under the interpretation. Thus, at least one interpretation of the names and predicates is possible, given the truth

[70] See also Button & Walsh (2018).

conditions of the sentences.

But if *one* interpretation of the names and predicates is possible, then so are many others. This can be proved using a permutation argument. The idea is to first take the interpretation *I* that maps names to objects and predicates to extensions from the domain D, thereby rendering each sentence of the theory true. Next we take an arbitrary permutation μ of the domain, a bijective function from D to D. Let's say that μ is not the identity function. We can then define a new interpretation *I** that is parasitic on *I* and μ. Names under *I** are mapped to their images of $\mu \circ I$. The predicates under *I** are stipulated to apply to an object x just in case it applies under *I* to y, and x is the image of y under μ. It is then provable that a sentence of the language is true-under-*I** if and only if it is true-under-*I.* Therefore *I* and *I** make all of the same sentences true, and so they generate the same truth conditions for each sentence in Paul's language. So if the interpretation of subsentential expressions is entirely determined by truth conditions, then *I* and *I** are equally good as far as our interpreter is concerned. But *I** is defined by an *arbitrary* permutation, so its reference scheme could be totally bizarre. For example, it may assign the predicates to randomly-assorted extensions, with no natural unity.

The main lesson to draw from this permutation argument is that sentential truth conditions vastly underdetermine the reference of words. This poses a threat to any theory that takes reference to be purely determined by an antecedent assignment of truth conditions, with no further constraints. It poses an immediate *prima facie* problem for Davidson, whose interpretation scheme would have us interpret words on the basis of a collection of T-sentences. It also poses an indirect *prima facie* problem for Lewis. Lewis' system assigns contents to sentences and these contents can be understood as a mapping from worlds to truth values. And even though these contents supply more fine-grained semantic information than Davidson's extensional semantic values, it is still possible to run a permutation argument to reach a similar conclusion about Lewis' intensional construal of semantic values. Simply define a permutation on the domain of each possible world and then define the parasitic interpretation scheme accordingly. So even for Lewis, the same lesson applies: truth-at-a-world does not uniquely determine reference-at-a-world.

Notice, however, that the permutation argument does not even present a *prima facie* problem for the functionalist. As a version of productivism, the view takes the referents of words to be fixed prior to sentential truth, and the truth conditions of sentences depend directly on the referents of their parts. Within this grounding structure, there can be no worry that the reference facts will be left unfixed by sentential truth.

4.5 Mitigating truth-to-reference indeterminacy

I said that the permutation argument poses a *prima facie* threat to Davidson and Lewis' versions of interpretationism. But the threat is decisive only insofar as there are no further constraints on reference determination, besides generating the right truth conditions. In fact, each author has further resources to draw from.

In this section, I will briefly overview the further constraints that allow for interpretationists to mitigate the problem of truth-to-reference indeterminacy. But before we attend to the details, we can describe each of their strategies abstractly, as follows. The key to blocking truth-to-reference radical indeterminacy, for the interpretationist, is to argue that not all interpretations are equally good, even when they assign the same truth conditions to sentences. Some interpretations must be better than others, for reasons independent of the sentential meanings they generate. To borrow a term from Lewis, let's call the potential extensions that are intrinsically fit to serve as subsentential meanings 'eligible'. The strategy, then, is to explain what makes a potential interpretation eligible. Each of our interpretationists provides different answers to this problem.

4.5.1 Davidsonian projectivism

I will start with Davidson. Unfortunately for us, Davidson does not specifically address the permutation argument in his writings. He often admits that interpreting a subject can result in multiple different, equally good semantic theories for their language (e.g. 1977: 256). But he does not address the possibility of the perverse interpretations provided by permutation. Thus, to develop the Davidsonian response to this kind of indeterminacy, we need to engage in some interpretative interpolation.

To this end, I draw upon Glüer (2018). In this paper, Glüer is particularly interested in analyzing the role that the interpreter plays in Davidson's metasemantics. She specifically addresses the question of where, if at all, the interpreter has a role to play in the determination of meaning.

To answer this, it is helpful to consider her own example of interpreting an alien (2018: 236). Like the previous thought experiment, we imagine that we are interpreting this new subject from the context of radical interpretation, with no previous knowledge of their meanings or beliefs. But unlike the previous thought experiment, this time we are to imagine that throughout our entire series of observations of utterances paired with circumstances, we are unable to discern any noticeable pattern. To *us*—the interpreters—there are no noticeable similarities among the things to which the alien applies their terms. Suppose they have a predicate and we observe them variously applying it to a tree, our shoe, a clock, an animal, and other random objects.[71] We do not notice any natural way to classify the objects to which they apply their term.

As Glüer points out, it is possible for us to assign an extension to this predicate while adhering to the principle of maximizing truth. We simply list off its extension by enumerating every case of application (236). We could, in fact, do this to every term and predicate in their language, provided that we are omniscient about every occasion of use and their linguistic dispositions. We would then end up with a semantic theory that renders everything that the alien says true. However, the result will assign massively heterogenous extensions to the terms, with no discernible unity.

Glüer then considers whether this haphazard semantic theory would count as conforming to the Davidsonian principle of charity. Her answer is that it would not. The reason for this is because it violates another constraint of Davidson's metasemantics. Namely, the haphazard semantic theory does not allow the *interpreter* to *understand* the alien (237). For a semantic theory to achieve this, it must do more than merely list off the extensions of each term in a way that appears arbitrary to the interpreter. There must also be some degree of shared conceptual resources between the interpreter and the subject, so that the interpreter can state the subject's meanings *in the interpreter's own terms*.

To justify this interpretation, Glüer observes that Davidson often hedges his statement of the principle of charity with a reference to the interpreter. As a case in point, he writes, "the Principle of Correspondence [an aspect of charity] prompts the interpreter *to take the speaker to be responding to the same features of the world that he (the interpreter) would be responding to under similar circumstances*" (Davidson 1991: 211). Given this principle, which is grounded in the connection between interpretation and understanding, we get the result that a charitable interpretation is one that seeks to maximize common conceptual ground between the speaker and interpreter. It must, as much as possible, seek to correlate the speaker's terms with the interpreter's own concepts (Glüer 2018: 235).[72]

We can put this point in terms of eligibility. For an interpretation of a language to be correct or best, it is not simply a matter of maximizing truth at any cost. It must also seek to maximize the eligibility of the properties assigned to the predicates. (Here we are thinking of properties as abundant; any set of objects corresponds to a property.) The supposition is that some properties are more eligible than others: they are more fit to be meanings and thus they

[71] Let's suppose, fancifully, that we had some reason for suspecting the term to be a predicate.

[72] Glüer writes, "To even come up with the hypothesis that Kurt [a fictional subject] is talking about rain when uttering 'es regnet', the world at those times needs to discriminatively display features that strike the interpreter as *similar*, features the interpreter either already has a concept of, or can form one for upon recognizing the similarity" (2018: 235).

make for better interpretations. On Glüer's reading of Davidson, 'eligibility' amounts to *detectable similarity, as recognizable to the interpreter* (245). When a collection of objects strikes the interpreter as inherently similar, then that collection is eligible to be the extension of a predicate in a charitable interpretation. But when a set of objects exhibits no inherent similarity that the interpreter can detect, it is thereby less eligible.

Supposing that this faithfully interprets Davidson's theory, we now have the materials for a Davidsonian response to the problem of truth-to-reference radical indeterminacy via the permutation argument. It is observed that the 'non-standard' interpretations that are parasitically defined by permuting the domain have been obtained from permuting the domain *arbitrarily.* It follows that the non-standard interpretations will assign extensions to the predicates that are utterly random. There is unlikely to be any detectable similarity between the objects assigned to the extensions of the predicates. But in that case, the Davidsonian interpreter (given that they speak a natural language like English) will not have any concepts to reasonably subsume the extensions assigned by the non-standard interpretations. And it is for this reason that the non-standard interpretations will not count as eligible by the lights of Davidson's principle of charity. In short, they won't allow the interpreter to understand the subject, and thus they don't compete with the standard interpretation for the best overall semantic theory.

While this conception of eligibility does appear to succeed in ruling out deviant interpretations and thus restoring a degree of determinacy, it also has an obvious shortcoming. The downside to this view is that it construes semantic determination as essentially perspectival. The meaning of a subject's words essentially depends on the perspective of a second-personal interpreter. In this scheme, it is *because* we, the interpreter, find more similarity among the items of class X than class Y, that it becomes the case that a speaker's words apply to the Xs rather than the Ys. This is so even when the speaker belongs to an alien speech community. Even then, the speaker's meanings will be hostage to the interpreter's conceptual scheme—despite the fact that the interpreter is a foreigner. This interpreter-relativism, or projectivism, about semantic properties strikes me as an implausible claim to make about the metaphysics of meaning and reference. Presumably, what a speaker means ought to depend solely on *them* and *their* speech community. This Davidson-inspired doctrine may very well escape the charge of radical indeterminacy, but the price is steep.

4.5.2 Lewisian Naturalism

We see that for Davidson, the interpreter has an indispensable role to play in the fixation of semantic properties. The interpreter can be relied upon to settle unwanted indeterminacy, but at the cost of introducing a subjectivist element into the metaphysics of meaning. Lewis, by contrast, aspires to rid the metaphysics of meaning of these subjectivist elements. Although he invokes the radical interpretation thought experiment, he regards it to be a mere dramatic device:

> To speak of a mighty knower, who uses his knowledge of these constraints to advance from omniscience about the physical facts P to omniscience about the other facts determined thereby, is *a way of dramatizing our problem*—safe enough, so long as we can take it or leave it alone. (1974: 334)

According to Lewis, the facts about what a speaker believes, desires, and means ought to be entirely determined by the physical facts about them as a physical system and the environment that they inhabit (ibid: 333–4). Moreover, Lewis is less willing to tolerate a significant remainder of semantic indeterminacy once the interpretation procedure is done. He says that some indeterminacy in the subsentential meanings is inevitable (342–3), but for the attitudes and sentential meanings it ought to be kept rather minimal; "*Credo:* if ever you prove to me that all the constraints we have yet found could permit two perfect solutions, differing otherwise than in the auxiliary apparatus of [the compositional semantic theory], then you will have proved that we have not yet found all the constraints" (343).

Lewis specifically addresses the permutation argument in his paper 'Putnam's Paradox'. In this paper, he develops his own famous response to the problem of truth-to-reference indeterminacy. And unlike the Davidsonian response, this one is thoroughly objectivist.

Before I outline the solution, there is one word of clarification. It is not Lewis' stated goal in this paper to defend the entirety of his interpretationist picture. Instead, he couches his response as a defence of 'global descriptivism'. This is the view that each of a speaker's terms (including singular terms and predicates) obtain their referent by a matter of best fit to their overall theory (1984: 224). A reference scheme that renders one's sentences true (as much as possible) is thus a worthy candidate to be selected as the correct reference scheme. This differs from his official view in that it takes linguistic truth, rather than mental content, to be primary, and it doesn't assign any metasemantic role to communal conventions. Although this is controversial, I will assume that Lewis' response to the problem of radical indeterminacy in this setting can be brought to bear to defend something in the vicinity of his interpretationism from the model-theoretic argument.[73]

To set the groundwork for his solution, Lewis begins with the claim that generating the right truth conditions for sentences cannot be the only constraint on the selection of referents and subsentential meanings. (This much agrees with the Davidsonian solution presented above.) Besides generating the right truth conditions, the correct reference scheme must also assign eligible referents to the names and predicates. For Lewis, an object or a set is eligible insofar as it is intrinsically fit to be a meaning. This means that, overall, the best interpretation of a language must maximize the joint constraints imposed by truth and eligibility (1984: 227). Oftentimes this will be a balancing act, as the two constraints may trade off against each other.

The next step for Lewis is to say what eligibility consists of. To motivate his answer, he observes, as we did above, that the extensions assigned by the interpretations obtained by permutation will inevitably be mixed bags of randomly assorted objects. In Lewis' words, they are not 'natural' groupings: the members of these collections will not exhibit much similarity amongst themselves along any dimension. According to Lewis, this is what makes these sets ineligible to be the meanings of predicates. He says:

> Among all the countless things and classes that there are, most are miscellaneous, gerrymandered, ill-demarcated. Only an elite minority are carved at the joints, so that their boundaries are established by objective sameness and difference in nature. Only these elite things and classes are eligible to serve as referents. (1984: 227)

Eligibility, for Lewis, is thus a matter of natural demarcation. An eligible set will be such that its members are naturally grouped together by some respect which makes them similar and distinguished from the non-members.

It is important to note that, for Lewis, whether a set of objects is natural is a thoroughly objective matter. It does not depend on whether they strike *us,* or an interpreter, as similar. This is where his solution differs from the Davidsonian one. He writes, "If I am looking in the right place for a saving constraint, then realism needs realism. That is: realism that recognises… objective sameness and difference, joints in the world, discriminatory classifications not of our

[73] I am thus ignoring one source of tension between the global descriptivism of Lewis (1984) and the conventionalist metasemantics of Lewis (1975): the idea that meanings are reference magnets is at odds with the idea that meanings are conventional. (Granted, Lewis 1975 only claims that *sentential* meanings are conventional, through the conventions of truthfulness and trust. He does not apply his analysis of conventions to explain subsentential meaning.) See Simchen (2017: 30–2). The Lewisian analysis of conventions requires that a conventional solution to a coordination problem—e.g. fixing the referent for a term—has an element of arbitrariness, i.e. that there are at least two equally good candidate meanings. Indeed, Sider (2013) proposes that a necessary condition for meaning being conventional is that there is *not* a single candidate meaning that is more intrinsically eligible than the rest (ch. 4). So if the Lewisian interpretationist appeals to reference magnetism in response to the problem of indeterminacy, they may have to forfeit the claim that subsentential meanings are fixed by conventions. That is a significant step away from the general outlook of Lewis (1975).

own making" (228). Lewis also allows for the naturalness of groupings to come in degrees (227). Some groupings will be perfectly natural: in his opinion, these will correspond to the properties of fundamental physics (228). Other groupings will be less natural, such as biological and social classifications. But all of these are still much more natural than the miscellaneous sets assigned by the permuted interpretation.

Lewis' solution to Putnam's argument for indeterminacy can be summarized like this. Within the context of global descriptivism, the correct reference scheme is the one that best interprets the subsentential expressions given the truth conditions assigned to sentences. The best interpretation is now seen as one that must trade off between two desiderata: it must, as much as possible, seek to render the subject's theory true, and it must also, as much as possible, maximize the eligibility of the referents. The eligibility of referents consists in natural similarity. Thus, the best interpretation will seek to assign naturally internally-similar collections as the extensions of predicates. Finally, this last constraint serves to rule out the unintended interpretations that are concocted by Putnam's argument. Since these interpretations are obtained by arbitrarily permuting the domain, they will not assign natural groupings to the predicates, and are ruled out for this reason.

In broad outline, Lewis' answer seems to be exactly what the interpretationist needs. They need an additional constraint to rule out the perverse interpretations provided by Putnam's argument, and rejecting them on the grounds of unnaturalness is an intuitive move to make. The burden for Lewis, however, is to make this idea of 'objective naturalness' more precise. It is one thing to say, on an intuitive level, that the green things are more naturally grouped together than the grue things (to use a familiar example). It is another thing to articulate a theory as to why this is so.

Lewis does, in fact, offer a positive account of the degrees of naturalness in his (1984). His idea is that 'naturalness' can be cashed out as a matter of *definability in terms of the perfectly natural properties* (228). The perfectly natural properties are the ones that figure into our fundamental physics: mass, charge, quark colour and flavour (228). From these, Lewis believes that it is possible to define all other possible collections. The key idea is that the relatively more natural collections will be relatively simpler to define. It will require less connectives and less operators to define a chemical collection from physical properties than it would take to define a biological collection; it would take less to define a biological collection than it would take to define a social collection; finally, it would take less to define a social collection than it would take to define the motley assortments assigned by permuted interpretations (228).[74] We thus achieve a syntactical criterion to measure degrees of naturalness.

[74] Lewis writes, "Indeed, physics discovers which things and classes are the most elite of all; but others are elite also, though to a lesser degree. The less elite are so because they are connected to the most elite by chains of definability. Long chains, by the time we reach the moderately elite classes of cats and pencils and puddles; but the chains required to reach the utterly ineligible would be far longer still." (1984: 228)

Although this solution is clever, it is not without difficulties. J. Robert Williams outlines one such liability in Lewis' definition in his (2007). His objection is quite sophisticated, and I can only give the briefest description here.

Williams' objection applies to any consistent global theory that is indifferent to the number of things that there are. Suppose that there is such a theory in natural language that we intuitively take to be about ordinary macroscopic objects. If we take such a theory, then we can add to it a clause that says that there are exactly *n* things (a finite number). Once we have done this, then a theorem from Henkin (1949, 1950) says that there is a model for that theory with a domain of size *n*. Moreover, the domain of the model can be anything; in particular, we can choose it to consist of the numbers 1 through *n*. This model would be unintended, since the global theory (we may presume) was not intended to be about numbers.

As Williams points out, the numerical interpretation has an upper bound on its level of complexity. We can specify each predicate by simply enumerating its extension. Doing so will result in a long, but finite, definition of all of the properties, which will have syntactic complexity (number of connectives) *M.* This is the limit on the 'degree of naturalness' for this unintended interpretation.

As for the intended interpretation, to measure its naturalness, we must define the extensions in terms of perfectly natural properties. However, for Lewis, the perfectly natural properties reside at the most fundamental microscopic level. For the actual world, they are presumed to be the properties of quarks.

Williams then describes a way in which a possible world could make for more complex intended interpretations than the unintended arithmetical model. If it is true that quarks are the most fundamental microphysical atoms, then this is a contingent fact. There could be worlds with further physical layers beneath the layer of quarks, which replicate our world 'from the quarks up' (2007: 390). If so, then their perfectly natural properties will be even more steps removed from the macrophysical objects that make up the intended interpretation of the theory. If a world has enough layers, then the definitions of the intended extensions could *exceed* the definitions of the unintended arithmetical extensions in their complexity. In that case, then according to Lewis' 'naturalness' constraint, the theory of the speakers in those worlds would determinately refer according to the unintended arithmetical reference scheme. But that is a bizarre result!

This shows that Lewis' naturalness constraint has objectionable consequences if 'naturalness' is explicitly defined as Lewis defines it. For one, it implies that our reference to macrophysical objects, like tables and chairs, does not supervene on the macrophysical structure of the world (Williams 2007: 392). We could end up failing to refer to tables and chairs if the microphysical structure of our world were sufficiently complex—we would be referring to numbers instead! But this kind of failure of reference to supervene on macro-level objects, properties, and relations is unprecedented; it is not at all like Putnam's twin-earth.[75] Another worry stems from the fact that it is a live epistemic option that our own world meets the level of complexity to entail this undesirable outcome. After all, the definitions of macrophysical properties in terms of the properties of quarks will already be massively complex. Who is to say that it will be less complex than the definitions of the arithmetical extensions provided by the unintended interpretation? In that case, it could actually turn out, according to Lewis' theory, that we are referring to numbers when we intend to talk about tables and chairs!

The lesson to draw from this is that Lewis' official theory of naturalness in terms of definitional length to microphysical properties is problematic if understood as the sole saving constraint against unintended interpretations. An interpretationist still needs more resources to rule out bizarre unintended interpretations of a language and fend off the threat of truth-to-reference indeterminacy.

4.6 Order of explanation and metasemantic holism

It is now worth summing up the dialectical situation and bringing out the different general strategies for combatting radical semantic indeterminacy. As we saw in §4.2, metasemantic interpretationism, in both its Davidsonian and Lewisian forms, is tied to two theses about metasemantic determination: (I) that sentential meanings are fixed prior to reference and (II) that the conditions that fix reference are global constraints that pertain to entire reference schemes for a language (metasemantic holism).

For our purposes, it's important to discern how the problem of (radical) truth-to-reference indeterminacy relates to these two theses. To be specific, the problem arises owing to the fact that a determinate assignment of (unstructured) meanings to sentences does not uniquely determinate an assignment of meanings to words. (This is the lesson of the model-theoretic argument.) Therefore, any view that takes subsentential meanings to be determined (at least in part) by (unstructured) sentential meanings runs the risk of leaving subsentential meanings undetermined. Insofar as interpretationism is committed to (I), it falls within the crosshairs of this problem.

On the other hand, it's worth reiterating that functionalism, as a version of productivism, rejects thesis (I). Functionalism takes sentential truth conditions (meanings) to be determined by

[75] Putnam's classic thought experiments show that reference to macro-level natural kinds and substances (e.g. water) fails to supervene on macro-level properties and relations precisely because the kinds and substances themselves are not solely individuated by their macro-level features; they are also individuated by their microphysical structure. Needless to say, this is quite different from the proposition that reference to macro-level objects fails to supervene on the macro-structure of the world because we could be referring to numbers if the microphysics were sufficiently complex.

the pre-determined facts of reference. Since the present version of the indeterminacy problem depends on commitment to (I), the functionalist is immune.

Moreover, notice how each response on behalf of the interpretationist is shaped by their commitment to thesis (II). The interpretationists each take reference to be fixed by the global properties of overall interpretation schemes for object languages. For this reason, they appeal to such global features as *maximizing understanding or shared conceptual resources* (Davidson) or *maximizing the naturalness of the predicate extensions* (Lewis). Neither of these virtues operates on individual reference relations in isolation. In order for them to fix a particular fact of reference, the entire language and assignment of referents must be taken into account.

In short, the radical indeterminacy worry arises from thesis (I) and the interpretationists' responses are constrained by thesis (II).

We have already seen some reasons to think that Davidson's and Lewis' choices for global reference-fixing constraints are undesirable or inadequate. But in fact, there may be a stronger reason to think that *no* metasemantic theory that's committed to (I) can save itself from indeterminacy, regardless of its theory of eligibility for predicate meanings.

The argument comes from Simchen (2017). In brief, the argument begins by considering the intended interpretation $\mathcal{I}$ of some first-order language. Simchen's immediate target is Lewis and so he supposes that the predicate meanings are chosen to maximize Lewisian naturalness. However, we can run the same argument by considering alternative eligibility constraints, such as the Davidsonian one whereby the predicate meanings are chosen to maximize the overlap in conceptual resources between the speaker and interpreter.

Once we have our intended interpretation, we can construct the following perverse interpretation $\mathcal{I}^*$ (Simchen 2017: 40–2). First, the predicates are assigned by $\mathcal{I}^*$ to the same extensions that they are assigned to by $\mathcal{I}$. This guarantees that $\mathcal{I}^*$ will meet the global constraints that define eligibility. Next, we consider an arbitrary, non-trivial permutation *f* on the domain. Following this, we assign referents to the singular terms according to the rule that $\mathcal{I}^*(t) = f(\mathcal{I}(t))$. Finally, in order to ensure that $\mathcal{I}$ and $\mathcal{I}^*$ generate the same truth conditions for the sentences of our language, we devise a non-standard definition of truth-in-a-model, called 'scrambled-truth-in-a-model': a sentence $\phi(t_1,...,t_n)$ is scrambled-true (in our model) if and only if $<f^{-1}(\mathcal{I}(t_1)),...f^{-1}(\mathcal{I}(t_n))>$ is in $\mathcal{I}(\phi)$ (41). (If the terms include variables then this is relativized to assignments.) In short, the predicates are assigned their intended meanings, the singular terms are assigned unintended meanings, and the semantic compositional rules are adjusted so that all of the sentences are associated with their intended coarse-grained truth conditions.

The possibility of this alternative construal of reference and truth-in-a-model presents another challenge for any view that adheres to (I) and takes reference to be determined by its contribution to generating the right truth conditions. It does not present a problem for the productivist or functionalist, since they take truth to depend on reference directly, and they take reference to be settled by prior causal relations (51). However, it does present a challenge to the interpretationist insofar as they adhere to (I). The challenge, for them, is to give some reason to think that the standard interpretation $\mathcal{I}$, with its standard construal of truth-in-a-model, is preferable over the non-standard $\mathcal{I}^*$, with its non-standard construal of truth-in-a-model, as a representation of the real semantic facts. This problem is acute precisely because $\mathcal{I}$ and $\mathcal{I}^*$ agree on the eligibility of the meanings assigned. And besides that, the most obvious response is unavailable to the interpretationist: they cannot say that the standard construal of truth-in-a-model is preferable to scrambled-truth-in-a-model owing to the direct dependence of truth-conditions on reference. This response is blocked precisely by their commitment to (I).[76] (Simchen also considers and replies to several other possible responses from the interpretationist, but it would take us too far astray to mention them all (45–53).)

At this point, it is natural to wonder whether the interpretationist can salvage their general outlook by dropping their commitment to (I). After all, there appear to be several distinct ideas embroiled in the view. On the one hand, the interpretationist claims that meaning is constituted by features that surround the activity of interpretation, whether by an actual interpreter or an ideal interpreter. This represents a kind of metasemantic *perspectivalism*, since the perspective of the interpreter is essential to determining the facts of meaning. On the other hand, there are theses (I) and (II), which are specific claims about how the activity of interpretation must proceed. According to these claims, the process of interpretation must first target sentences and it operates holistically. One might wonder whether it is possible to develop a version of interpretationism that accepts the perspectivalist metaphysics but rejects the idea that sentences must be interpreted before subsentential expressions. (Ball 2017 suggests this in response to Simchen's argument.) In that case, the fact that my partner's use of '*that*' refers to *o* is seen as emerging out of the interpretability of '*that*' as '*o*' in the interpreter's metalanguage, but the factors that favour this interpretation will have to do with how my partner uses this term. Indeed, we might even say that this interpretation is favoured as a result of the causal relations between my partner's use of '*that*' and *o*. We might say that it is these reasons, rather than maximizing truth, that ground the fact that '*that*' *refers to o* from within the interpreter's point of view. This upholds the basic metaphysical picture of meaning as interpretability but it rejects the truth-first approach of Davidson and Lewis.

Of course, Davidson and Lewis have their own independent reasons for tying their

[76] One might wonder whether any analogous worry can be raised against the productivist, who takes reference prior to truth. Just as this argument *grants* to the interpretationist the standard assignment of truth conditions and then scrambles reference, perhaps we can task the productivist to rule out a semantic theory that assigns standard reference to subsentential expressions and then scrambles truth. But as Simchen argues (2017: 52), the productivist can rule out the scrambled construal of truth-in-a-model precisely because they envision the dependence of truth conditions on reference to be *direct.* In the productivist picture, the truth conditions of sentences are directly explained by the referential contributions of their subsentential parts—but not so for the interpretationist. So there is an asymmetry between the two positions which makes the interpretationist uniquely vulnerable to this kind of problem.

versions of interpretationism to thesis (I). For Davidson, it is because sentences afford the first foothold of empirical evidence into what an alien speaker means. And for Lewis, it is because sentences are the subjects of the conventions of truthfulness and trust. So it is not altogether clear that an alternative version of interpretationism without (I) is viable. However, it would take us too far afield to explore this issue further.

Instead of considering the possibility of interpretationism without (I), I propose instead to close this chapter by raising an alternative indeterminacy problem for interpretationism. Unlike the previous challenge, this one will target any version of interpretationism that is committed to (II)—metasemantic holism. This challenge will thus remain even if the interpretationist can unfasten their overall metaphysics from (I) and avoid the previous objection from radical indeterminacy. It will present a problem so long as they remain committed to (II). And as I remarked earlier, there is strong reason to think that the metaphysics of interpretationism is tied to (II). An interpretational theory is an empirical theory after all, and therefore, it must be confirmed holistically. So if, like the interpretationist, one holds that the determinants of meaning are essentially the same as the factors that confirm interpretational theories, one must subscribe to metasemantic holism.

The problem for metasemantic holism, as I see it, is that it gives an implausible picture of expressions whose metasemantics is sensitive to particular features of their context of utterance. Among the context-sensitive expressions, it is standard (following Kaplan 1989b) to distinguish between pure indexicals and supplementatives. (The term 'supplementatives' comes from King 2014b.) Pure indexicals, such as 'I' and 'now', are those whose surrounding linguistic conventions supply a context-invariant rule that determines their semantic content in any context. Supplementatives, by contrast, are not supplied with any such context-invariant rule; rather, they must be *supplemented* in some way to achieve a determinate semantic value when used in context. The paradigmatic examples are demonstratives like 'that'. In order for the expression 'that' to obtain a referent in context, the speaker must *do something* to exploit the features of their environment to pick out a particular object.

It is too crude of a picture to think that demonstratives have any standard *demonstration,* like pointing towards an object, that single-handedly fixes their referents in all of their various uses. For one, many uses of demonstratives don't require the speaker to make any physical gestures. For another, even when one does make a gesture, it will not be sufficient by itself to determine any particular object as the referent. To paraphrase Kaplan (1978), when one points towards a man, one also points towards his jacket, his shirt, his buttons, etc. For these reasons, we should not expect there to be a reference-fixing rule for demonstratives that is exhibited by the overt behaviour of speakers. There is no common behavioural pattern. However speakers make their referential intentions known, it is by exploiting particular features of their conversational context and environment to make their intended referent salient.

This raises a second point. It is eminently plausible to think that speaker intentions play a crucial role in fixing the referents of demonstrative expressions. Indeed, in light of the previous remarks, any view that doesn't recognize a role for speaker intentions will be hopeless. When my partner utters '*that* is a lighthouse', her demonstrative refers to *o* precisely because it was produced with the intention to refer to *o* in mind. This is also an example of a perceptual demonstrative. Thus, the metasemantic explanation of the content of the referential intention will advert to the content of her antecedent perceptual experience.

I take these remarks about demonstrative terms to be fairly commonplace. The point of raising them, however, is to show how they don't sit well with interpretationism and its commitment to metasemantic holism.

Allow me to borrow an example from Reimer (1991). Suppose that you and I are near a dog park. There are dozens of dogs running around, making it impossible for you to tell which one I'm focused on. I say, without any overt signals such as gesturing or pointing, 'that dog belongs to my neighbour'. There is a particular dog that I had in mind, however, nothing in my behaviour makes the dog salient to you. Hence, there is no way for you, my audience, to interpret my utterance as referring to any particular dog (short of some sort of brain scan that allows you to uncover the precise direction of my focus).

Now, there are two different ways of reading the situation. On the one hand, we might say that my utterance failed to express any proposition, by dint of the indecipherable use of 'that'. By failing to make salient which dog I intended to refer to, I therefore fail to secure any referent for my expression. If that's the case, then my utterance is neither true nor false, for there is nothing that I said. According to this way of reading things, we locate the defect in our conversation in the *semantics* of my expressions; the defect is that I uttered a referent-less expression. On the other hand, we might instead say that I *did* express a proposition—I said something true or false —but I failed to make it plain *which* proposition I expressed. In other words, my words really do mean something, but it is unclear to you what they mean. According to this way of reading things, the defect in our conversation is located in your lack of understanding, not in the semantics of my utterance. My utterance has determinate semantic properties, some of which are beyond your ken. What has gone wrong is a matter of epistemology; you don't *know* what I mean, but that isn't to say that my expression is meaningless.

It seems to me that the most intuitive response is the second option.[77] I really did mean something; you just couldn't tell what I meant. Moreover, what I said was *true* if and only if the dog I intended to refer to—and did in fact refer to—belongs to my neighbour. This response also

[77] King (2014b) takes the opposite conclusion and develops his metasemantics of demonstratives to include *interpretability by the audience* as a necessary condition. However, he offers his verdict on such cases as a brute intuition. In response, I take the phenomena of anaphoric reference on the basis of such demonstratives as offering a strong presumptive case in favour of the view that they succeed in referring, even in the absence of audience interpretability.

seems to be borne out by the linguistic data. You may respond with ‘which dog is your neighbour’s?’, presupposing that there is one. Moreover, you could rightly infer that there is *some* dog there that is my neighbour’s. However, the existential claim only follows on the assumption that I had asserted *some* singular proposition.

Finally, for perhaps the most compelling piece of evidence, notice that you, my audience, can successfully refer to the particular dog using anaphora. You might ask ‘does your neighbour treat him well?’ and this will have a determinate answer. Your pronoun ‘him’ inherits its referent from my demonstrative ‘that’, proving that my demonstrative had the referent that I intended, regardless of your inability to interpret it.[78] This just goes to show that there is a determinate fact about what my utterance referred to. My utterance referred to one dog, let’s call him Fido, and it did not refer to any of the other nearby dogs.

Granting that there is this semantic fact, it must have a metasemantic explanation. I worry that the interpretationist beholden to metasemantic holism will be incapable of providing one.

Consider how the interpretationist must attempt to formulate an explanation. Let’s say that it’s a determinate fact that my expression ‘that dog’ refers to Fido rather than Clover (another dog that happens to be nearby and within my field of vision). To make things especially tough for the interpretationist, let’s further suppose that Clover also happens to belong to my neighbour and looks indistinguishable from Fido. How will the interpretationist approach this question?

Pursuant to their holism, they must consider two competing interpretations for my entire idiolect. One interpretation (the correct one) $\mathcal{I}$ will assign referents to all of my singular terms, including, among other things, *Fido* to my present use of that ‘*that dog*’. Another interpretation (an incorrect one) $\mathcal{I}^*$ will assign referents to all of my singular terms, including *Clover* to my present use of ‘*that dog*’. Finally, for $\mathcal{I}$ to be favoured over $\mathcal{I}^*$ (and thereby determine its correctness, if all other things are equal), it must beat out $\mathcal{I}^*$ when we tally up their global virtues. Perhaps $\mathcal{I}$ is more charitable than $\mathcal{I}^*$ (renders true more of my assertions and beliefs), or better maximizes naturalness of assigned extensions, or better maximizes shared conceptual resources between speaker and interpreter.

The basic problem for interpretationism, as I see it, is that exclusive focus on the global virtues of wholesale interpretive theories leaves a metasemantic view unlikely to single out a referent in this case. Right away, we observe that $\mathcal{I}$ and $\mathcal{I}^*$ may be on par as far as eligibility is concerned. They both assign precisely the same extensions (or intensions) to the predicates of my idiolect, and so they are equal on the score of Lewisian naturalness and Davidsonian eligibility. They may also be equally simple and predictive. Moreover, they may even be equally charitable.

[78] The alternative is to say that the speaker’s anaphoric pronoun also fails to refer. This option is unattractive because it does not explain the speaker’s behaviour. The speaker may utter ‘does your neighbour treat him well?’ under the impression that they successfully referred, even if they cannot tell which dog I referred to.

There may be no proposition that I believe of Fido that is not also true of Clover. Or if, perchance, there are things that I believe that are true of Fido and false of Clover and pertain to this specific encounter (e.g. <Fido is currently running past the tree>), they may be offset by other beliefs of mine where I have mistaken Fido for Clover (e.g. I believed <Fido is barking> when in fact Clover is Barking). The point is, it is possible for my belief set to be such that $\mathcal{I}$ and $\mathcal{I}^*$ come out equal on charitability.

Finally, since I did not make any distinctive gestures towards the referent, the interpretationist cannot appeal to any *ad hoc* interpretational rule for assigning referents on the basis of overt speaker behaviour. Generally speaking, the interpretationist is free to include in their guidelines for interpretation rules of the sort:

> if the speaker uses a demonstrative *D* accompanied by a gesture of type *G* towards *x*, then favour the interpretation that assigns *x* to *D*.

But such rules would be of no help for this example. From the audience or interpreter's point of view, there is nothing in my overt behaviour that favours one referent over another. *A fortiori*, there cannot be any general interpretative rules for assigning a referent to my use of 'that dog', short of investigating my particular cognitive relations to Fido.

All of this points to an inevitable lesson: to interpret my use of 'that dog', we need to heed the specifics of this case. Specifically, we need to heed my referential intentions, whose content is determined by the particular cognitive relations that I stand to the referent. It is unlikely that this can be done by citing any general constraints on what makes for the best overall interpretation—that is, unless we include in those constraints a sensitivity to the cognitive facts about the speaker. This is for two reasons. For one, the cognitive facts about me, the speaker, are determined by facts that are specific to my context and environment and will therefore elude any theory beholden to metasemantic holism. For another, it is a plain fact about this case that the second-personal point of view (whether that be you, my audience, or another second-personal interpreter) is ill-equipped to decipher the referent of my demonstrative. The case was designed so that facts that are privy to the interpreter will underdetermine the best interpretation. This would be so even if the interpreter was 'ideal' in the sense that they know *all* of the surface-level physical facts about the speaker, short of the cognitive facts.

Now, I do not take any of this to add up to anything like a knock-down argument against interpretationism. I take this point to show that the interpretationist's particular kind of holism is implausible and that it leads to a problematic case of semantic indeterminacy. But this is unlikely, by itself, to rouse a committed interpretationist. In fact, it may be possible for the interpretationist to amend their view to accommodate it.

How might the interpretationist adapt their view to explain the apparent facts of this case?

Evidently, they must take the determinants of the best interpretation to include the cognitive facts about the speaker, including the content of the speaker's perceptual states. Moreover, this content cannot be conceived as coarse-grained truth conditions or sets of worlds; it must be fine-grained enough to specify particular objects. In effect, the interpretationist must say that $\mathcal{I}$ is favoured over $\mathcal{I}^*$ in virtue of $\mathcal{I}$ assigning the object of my referential intention to my token of 'that dog', and that the content of my intention was determined by my perceptual state at the time of production. This isn't a global or holistic virtue of $\mathcal{I}$; rather, it's just to say that $\mathcal{I}$ corresponds to the cognitive facts of the case.[79]

Can the interpretationist say this? Perhaps. We had just noted that interpretationism appears to have two distinct dimensions. One dimension is the broad metaphysical claim that meaning is a matter of interpretability—that meaning is determined by the factors that guide interpretation. The other dimension consists of theses (I) and (II), which are two substantive constraints on how interpretation works and how metasemantic explanations are structured. We also noted that these two dimensions appear to be logically separable.

The present suggestion, in light of the foregoing problems for securing demonstrative reference, is to keep the claim that reference is grounded in interpretation, but ditch *both* of the substantive constraints on how interpretation works. In that case, this *unprincipled* kind of interpretationism would say that my token 'that dog' means Fido *because* it is interpretable as meaning Fido. However, its interpretability is not based on truth-maximization or any other holistic consideration like charitability. In fact, it is not based on any substantive principles derived from reflection on the context of radical interpretation. Rather, it is based on the specifics of my cognitive situation and the causal history of the term. This interpretationist keeps to the spirit of interpretationism only insofar as they agree that *interpretation precedes reference* in the order of metaphysical explanation. But other than that, there is little else in common with the views of Davidson and Lewis.

Even though there may be conceptual space for such a view, it is hard not to get the impression that the notion of interpretation would no longer be carrying any explanatory weight. The more that the interpretationist mimics the productivist, the more redundant their framework becomes. It's true that someone *could* say that the meaning of a term is grounded in its interpretability, and that the interpretability of a term is determined by its cognitive and causal history. But then I wonder why we don't just cut the middle step, since it doesn't appear to be

[79] Williams (2020) is an example of a recent, comprehensive version of Lewisian interpretationism. Strikingly, his theory exhibits the pattern that I am speaking of here: it cedes territory to productivism by relying on an assignment of contents to perceptual states that are fixed by their circumstances of production.

doing any work.[80]

The classical views of Davidson and Lewis didn't face this problem, since, for them, the notion of interpretation has teeth. Within their respective theories, general considerations about the process of interpretation yield substantive constraints on how meanings get determined. Specifically, it yields the structural constraints (I) and (II), as well as other specific claims that fill out the details (their precise versions of charitability, eligibility, etc.). But with the structural constraints gone, it's hard to see what's left of the claim that interpretation determines meanings.

So even though this last indeterminacy problem may not pose a decisive objection to interpretationism, it does leave them with a dilemma. The fact is, the metasemantics of perception-based demonstratives will be hopeless if it ignores the real cognitive facts about the speaker, some of which may not be privy to the second-personal observer. In light of this, the interpretationist has two options.[81] They can nevertheless insist that the methods of interpretation substantially constrain our metasemantics, which gives us (I) or (II) (or both). In that case, the notion of interpretation does real explanatory work within the theory. However, this option also raises the spectre of semantic underdetermination; it makes it likely that certain demonstrative tokens, like my token of 'that dog', are uninterpretable, and hence lack a determinate referent. Alternatively, the interpretationist can choose to make their metasemantics less constrained. In that case, the interpretationist may even allow for the correct interpretation to be determined by the cognitive facts of the speaker and the causal histories of their referring tokens. The problem, however, is that this renders the interpretationist framework redundant. Since the notion of interpretation is no longer doing any real work, it becomes a dispensable part of the view.

4.7 Conclusion

The broadest aim of this chapter has ultimately been to defend the functionalist account of reference against a range of worries and objections that surround one of its main rivals, metasemantic interpretationism. To this end, it was important to show how the functionalist view contrasts with interpretationism, both with regards to its claims about reference determination and its claims about the nature of the reference relation. One reason that this is especially

[80] Notice the parallel between this point and the objection raised against deflationism is chapter two. There it was argued that the deflationist must ultimately mimic the inflationist (particularly, the productivist) in order to account for the facts surrounding singular reference. Once it's admitted that the deflationist must engage in this mimicry, that zaps away the motivation for their view.

[81] That is, provided we continue to assume that some facts surrounding demonstrative use are within the scope of *semantics*. Perhaps an interpretationist could insist that this phenomena ought to be relegated to pragmatics, and hence brought outside the scope of their theory. I cannot comment much on this option here, other than that it seems like an *ad hoc* curtailment of their theory to avoid an otherwise serious problem.

important is because the interpretationist faces a presumptive challenge in securing an intuitive level of referential determinacy. However, as we have seen, the problem for them arises out of features that are distinctive to their view; the functionalist, by contrast, is immune.

We also considered several ways in which the interpretationist may respond to this challenge. But ultimately, I argued, there remains a reason to think that their view lacks security for the special case of perceptual demonstratives—that is, so long as they adhere to the doctrine I called *metasemantic holism.* The basic problem for them is that it is downright implausible to think that the fixation of perceptual demonstratives will be hostage to the global features that recommend overall reference schemes for a language. It is much more plausible to think that reference for perceptual demonstratives is fixed by features that are specific to the contexts in which these terms are produced, such as speaker intentions and other cognitive states. Any view that ignores these will run the risk of semantic indeterminacy. And any view that incorporates these will be a far cry away from the interpretationist's original idea that meaning is determined by interpretation.

Chapter 5: The Roles of Truth for Inquiry About the World

It is plausible to think that a theory of truth ought to have some bearing on the big thematic questions concerning first-order inquiry. (By ‘first-order’, I mean any topic concerning worldly objects and their properties—not our representations thereof.) After all, a theory of truth is essentially a theory of the connections between language, thought, and the world. And first-order inquiry primarily consists of interrogating questions about the world posed in a medium of language and thought. So, taken together, how could the two not be related?

Nonetheless, there are a couple of trends that make the connection between theories of truth and the concerns of first-order inquiry more obscure. For one, the metaphilosophical picture that sees certain first-order questions—i.e. the ones studied in philosophy—as answerable by linguistic and conceptual analysis has been on a steady decline since the mid-twentieth century. This isn’t to say that it doesn’t still have adherents, but they now form a heterodoxy. The current dominant view is that typical philosophical questions (specifically, those that aren’t ostensibly about meanings or concepts) are not essentially different from ordinary scientific questions insofar as they cannot be decided on semantic considerations alone (see Williamson 2007 and Taylor 2019).

The other trend is the rise of deflationist theories of truth and the ‘problem of creeping minimalism’ that they invite (Dreier 2004). In short, the deflationist theories have a habit of appropriating and mimicking the claims of their inflationary opponents. But the more that deflationists sound like inflationists, the more difficult it becomes to distinguish their implications towards other matters. Given this trend of appropriation and mimicry, it might appear that little else could hang on the debate between them.

This chapter argues that the theory of truth isn’t entirely indifferent to the concerns of first-order inquiry, despite these two trends. Deflationary and inflationary accounts of truth have their own distinctive consequences for our modes of investigation into the nature of things, including things that aren’t ostensibly related to semantics. To argue this, I will trace out the applications of the truth concept that are permissible according to deflationism and inflationism. Unsurprisingly, we will find that the differences will matter most when the topic of inquiry is distinctly philosophical. There is a perennial concern in philosophy to understand how reflection on semantic notions—i.e. truth, reference, representation—can be brought to bear on our investigation of the world. This broad question will be the overarching concern of this chapter.

A couple of examples may serve to narrow our topic. Suppose, for the sake of illustration, that we are engaged in an inquiry into one of the innumerable first-order questions of philosophical interest. Perhaps we are concerned to know whether a certain substance bears a certain property *essentially*—e.g. whether water is essentially H_20. Or perhaps we are concerned about whether a given action is *morally wrong*—e.g. whether factory farming is wrong. Whatever our question, imagine that we have managed to formulate it in the clearest possible terms. Given this much, a further question arises: how might the concept of truth figure into our inquiry? What intellectual advantage, for the purpose of investigation into our chosen topic, is conferred on us by having the concept of truth at our disposal? What might we achieve that we couldn't have done without thinking in terms of truth? In short: *what role does the concept of truth play in inquiry about the world?*

5.1 Deflated roles for truth

To approach this question, it is best to start with the view that assigns the most minimal role to the concept of truth: the deflationary theory, developed by Quine (1986), Field (1994a), Leeds (1995), and Horwich (1998a). Although the views of these authors differ in important ways, there is enough in common to distill the distinctly deflationary roles for the concept of truth.

Deflationism is commonly glossed as the view that truth doesn't require a 'deep' account. The reason for this, according to deflationism, is that truth can be entirely captured by a list of trivialities: 'snow is white' being *true* is simply a matter of *snow being white*; 'grass is green' being *true* is simply a matter of *grass being green*; and so on. Apart from this, there isn't anything more to say to explain the truth of these sentences or the propositions or thoughts expressed. In particular, we don't need an additional account of the relations between our representations and the world to explain the nature of truth.[82] For the deflationist, truth does not have the kind of nature that requires a deep investigation to reveal.

Thus described, deflationism is a negative metaphysical thesis about the *nature* of truth. It says that the aforementioned trivialities, taken together, entirely capture the property of truth. However, deflationists also justify this thesis with another set of claims about the *concept* of truth (or truth predicate). Specifically, they claim that the truth concept's primary function is to fulfill certain logical or syntactic needs and expedite our means of expression. As Quine (1986) and Leeds (1978) observed, the truth concept (or predicate) affords a means of expressing a plurality of statements without having to express each one. Suppose that we wished to assert a large (perhaps infinite) number of things about a given topic ($S_1, S_2, S_3,...$). The truth concept allows us to assert that *every one of those things is true* ($\forall$x: x$\in\{S_1, S_2, S_3,...\} \rightarrow$ x is true). Without the concept of truth, we would be incapable of formulating this succinct assertion using an ordinary quantifier.

[82] This is a first-pass description of their view. In order to handle certain difficulties, the deflationist has to add to their stock of resources. See chapter two.

We can gloss this role by saying that the truth concept is a device for expressing large (or infinite) conjunctions and disjunctions. Deflationists typically take the hardline stance that this is the sole reason for having a truth concept in our repertoire (and truth predicate in our language): it is just for the sake of increasing our expressive power (Quine 1986: 11–2; Horwich 1998a: 2–5). Indeed, a common strategy for justifying their view is to argue that each putative use of the truth concept can be uncovered as a covert generalization over statements which don't essentially mention truth.[83]

In order for the truth concept to perform this expressive role, it must be that a statement *S* is equivalent (in some sense) to the statement *'S' is true.* Thus *snow is white* must be equivalent to *'snow is white' is true*; *grass is green* must be equivalent to *'grass is green' is true*; and so on. In short, every instance of the disquotational schema must hold (see §2.2):

(DS) 'S' is true if and only if S.[84]

These equivalencies are key to moving between a generalization of the form $\forall x\text{: } x \in \{S_1, S_2, S_3, \ldots\} \rightarrow x \text{ is true}$ (e.g. *everything said in Newton's Principia is true*) and outright statements of the instances (e.g. *to every action there is always an opposed and equal reaction*). Indeed, for the deflationist, the explanation as to *why* the instances of DS hold is that they are engendered by the logic of 'true', which, in turn, is explained by its role in expressing generalizations.

Since a statement or thought is (in some sense) equivalent to an ascription of truth to it, it follows that the concept of truth facilitates another vital function that will become central to our concerns here: it is our primary device of semantic ascent and descent. Semantic ascent is standardly understood as the transition between the *use* of a sentence or thought—e.g. *water is* H_20—and a higher-order sentence or thought that *mentions* the sentence or thought previously used—e.g. *'water is* H_20*' is true*. We thereby go from speaking or thinking about things and their properties to speaking or thinking about *our representations* of those things and properties—e.g. from thinking about *water* to thinking about the *concept* of water. Semantic descent goes the other way. In semantic descent, we go from speaking or thinking about thoughts or sentences—e.g. *'water is* H_20*' is true*—to using the sentence or thought previously mentioned—e.g. *water is* H_20.

We can now recast the question of our initial inquiry. The concept of truth is our central tool for semantic ascent and descent. Therefore, the question of how truth is used in rational inquiry is tantamount to the question of how *semantic ascent and descent* are used in rational inquiry. We may now ask: when pondering first-order questions about the nature of things, why

[83] Armour-Garb & Beall (2005): 12; see Horwich (1998a) for the strategy; and see Gupta (1993) for a critique.

[84] A properly fleshed-out version of deflationism will restrict the schema to an appropriate set of sentences or propositions. But we need not concern ourselves with these technicalities here; see §2.2.

might we want to ascend and examine our representations of those things? In thinking, for example, about whether *water is essentially H_20*, for what purpose might we want to consider our concept of water? Or, in questioning whether factory farming is wrong, for what purpose might we want to consider our concept of *wrongness*?

Returning to deflationism, it appears that we've already found their answer. For them, semantic ascent and descent allow us to trade between individual uses of statements and large conjunctions or disjunctions of them. This idea was famously expressed by Quine;

> Where the truth predicate has its utility is in just those places where, though still concerned with reality, we are impelled by certain technical complications to mention sentences. Here the truth predicate serves, as it were, to point through the sentence to reality; it serves as a reminder that though sentences are mentioned, reality is still the whole point. (1986: 11)

Suppose that we are investigating the nature of water. For the deflationist, the primary reason why we might want to invoke truth and semantically ascend is that we may need to succinctly express a large body of chemical theory.

At this point, it is helpful to follow Rattan (2010, 2016) and distinguish between *explicit* and *inexplicit* truth attributions. Let's say that a truth attribution of any kind is a thought or sentence that predicates truth upon a truth bearer. We will call a truth attribution *explicit* when it attributes truth to a sentence or thought that is expressed in a highly transparent way. For the linguistic case, let's further stipulate that the subject making the attribution understands the object sentence. For example,

- 'Water is H_20' is true
- 'Factory farming is wrong' is true

are both explicit truth attributions. A distinctive feature of explicit truth attribution is that a subject who essays one is thereby in a position to use the sentence or entertain the thought to which they're attributing truth.[85]

Contrast this with *inexplicit* truth attributions, where the relevant truth bearer (or bearers) is not explicitly represented. Examples of this kind include:

[85] If we assume a Davidsonian (1979) analysis of quotation, then, for the sentential case, explicit truth attributions have another distinguishing hallmark: the sentence enclosed in quotation marks is both *mentioned* (as an object referred to) and *used* (as a demonstration of the mentioned sentence). Moreover, for Rattan, who's primarily interested in mental truth bearers, the hallmark of explicit truth attributions is that a subject who essays one is thereby entertaining the very thought to which they attribute truth (2016: 233).

- What Newton wrote in his *Principia* is true.
- Everything that the Pope said is true.
- Darwin's theory of evolution is true.

Each of these has the feature that it does not explicitly express the content of the sentences or thoughts that are represented as true. A subject who makes this kind of truth attribution may not be in a position to use the sentences or entertain the thoughts to which they attribute truth. Indeed, they do not even need to know *which* contents are being called true (Rattan 2016: 233).

According to Rattan (2016), the deflationary theory entails that the 'cognitive value' of the concept of truth resides solely in its use in inexplicit truth attributions (244). (The use of a concept has cognitive value if it makes some valuable epistemic contribution to our cognitive lives (2010: 139, 141).) This is because inexplicit truth attributions are the only kind that requires truth to perform its characteristic role in expressing infinite conjunctions. ('*Everything that the Pope said is true*' may be cashed out as '*(if the Pope said P_1 then P_1) and (if the Pope said P_2 then P_2) and ...*', where each sentence P_i is used in at least one conjunct.) On the other hand, according to Rattan, the deflationist ascribes the same cognitive value to the higher-order representation *'S' is true* as they do to the first-order representation *S.* For this reason, they cannot allow explicit truth attributions to play any significant role in our cognitive lives. Rattan further argues against deflationism by challenging this claim—that explicit truth attributions have no cognitive value. We will return to Rattan's objections shortly.

If, indeed, the deflationist claims that explicit truth attributions have no vital role in rational inquiry, then it's natural to think, at first, that their position is the intuitive one. It certainly seems unlikely that thinking about the *truth* of our thoughts or sentences will give us any greater purchase on how things stand in the world—as opposed to simply experiencing the world and thinking about it. To investigate, say, whether there are any orangutans left in Borneo, it would be of no help to reformulate our question as whether *the sentence* 'there are orangutans left in Borneo' *is true.* If we want to know about the orangutans, we should just go to Borneo and look. The point is, there ought to be an initial presumption against the claim that semantic ascent is a fruitful method for inquiry when the objects of concern are neither thought nor language. The philosopher who insists that there is such a role for semantic ascent is the one who bears the burden of proof.

In the following sections, I will canvas several non-deflationary uses of explicit truth attributions that tell against this presumption. But before I do, I would like to point out that, in my view, the deflationist can also recognize more roles for semantic ascent and explicit truth attributions, in addition to the core role that they ascribe to truth in inexplicit truth attributions. This goes against a prevailing opinion in the literature that deflationism can *only* recognize the utility of truth in inexplicit truth attributions (cf. Williams 1999; Rattan 2016). It is true that deflationists often claim that the *raison d'être* of the truth concept is to express generalizations. But even if the concept evolved to perform this function (so to speak), it doesn't follow that it can't be exapted for other purposes. Ascribing the truth concept additional purposes may be consistent with deflationism. In my view, it depends on whether truth itself remains deflated. For the deflationist, mentioning truth should be acceptable provided that nothing more is presupposed about it than what is captured by the instances of DS. The inflationist, on the other hand, may recognize mentions of truth that presuppose more. The distinction between deflationary and inflationary uses of truth is therefore *not* a matter of whether semantic ascent strictly serves no purpose other than covert generalization. Rather, it's whether truth is being used as a *mere* device of semantic ascent, or whether it's being used to report on substantial language-world relations.

The main reasons for recognizing further *deflationary* roles for explicit truth attributions stem from Quine. When it comes to truth, Quine is an archetypal deflationist. For him, the truth predicate really is nothing more than a device for switching between using and mentioning a sentence, and the main cause for doing so is the aforementioned 'technical complications' concerning quantification (1986: 10–3). However, Quine also famously advocated for semantic ascent as a fruitful method for philosophy, which for him, is not any different from ordinary scientific inquiry. And not all of the Quinean uses of semantic ascent cast truth in its standard role as a device of generalization.

For instance, according to one familiar Quinean theme, semantic ascent delivers a venue where we can ponder the truth of theories while neutralizing their ontological commitments. To use the well-worn example from his (1948), rather than deliberating over whether *Pegasus exists*, we can instead rephrase the question as whether *the sentence* 'Pegasus exists' *is true.* The rephrased question has the advantage of only mentioning the sentence 'Pegasus exists' and not mentioning Pegasus. Quine writes, "in so far as our basic controversy over ontology can be translated upward into a semantical controversy about words and what to do with them, the collapse of the controversy into question-begging may be delayed" (1948: 35).[86]

Notice that this transition takes an explicit truth attribution, not an inexplicit one. So this isn't a use of the truth predicate (or concept) that deflationists often advertise. Nonetheless, I see no reason why they can't accept it as important or legitimate. After all, it doesn't make any demands on the nature of truth besides what the deflationist allows. It's entirely consistent with the claim that there's nothing more to the truth of a sentence than what's given by its disquotational truth condition.

The same could be said for another, closely related, Quinean theme. Semantic ascent affords us the opportunity to *paraphrase* or *reconceptualize* our sentences or thoughts in ways that make them more advantageous for inquiry. It might not be helpful to phrase our initial

[86] The same point is repeated in his (1960), p. 272.

question as whether *Pegasus exists,* since mentioning Pegasus presupposes its existence. However, we can semantically ascend and exchange the sentence for something more suitable, e.g. ‘there is a winged horse caught by Bellerophon’. We then proceed to investigate whether *there is a winged horse caught by Bellerophon*.[87]

This paraphrasing maneuver requires a device of semantic ascent and descent, along with a story about how we choose the appropriate translation. Deflationist truth is perfectly suited for the former task. And as for the second, some accounts of translation and paraphrase are friendly to deflationism while others are hostile to it. It depends on whether we take facts of the sort *S_1 translates/paraphrases S_2* (or *S_1 has the same meaning as S_2*) as grounded in real representation relations between sentences and their subject matter (see §2.3). If, like Quine, we take paraphrase and translation to be guided by pragmatic considerations and unbeholden to any predetermined facts of reference, then this *ascend-and-paraphrase* tactic is entirely compatible with deflationism.

There’s one more Quinean role for truth that deserves our consideration. Within Quinean philosophical methodology, the truth concept facilitates semantic ascent for the purpose of viewing the holistic virtues of theories. Insofar as theory choice is guided by these holistic virtues, this gives the truth concept a distinctive role in our decisions as to which first-order propositions to assent to, and so ultimately what we think about the world. Suppose, for instance, that we are investigating the solar system at the time of Copernicus and we are specifically concerned with knowing whether the earth revolves around the sun. We may reason as follows: the earth revolves around the sun if and only if ‘the earth revolves around the sun’ is *true*; ‘the earth revolves around the sun’ is true if and only if its surrounding theory *T* is true; *T* is simpler than its geocentric competitors; therefore, *T* is true; therefore, the earth revolves around the sun.[88]

Once again, this is a use of the truth concept that performs an essential role in informing our view of the world (e.g. the solar system). It also involves an explicit truth attribution. But, as far as I can tell, it is entirely congenial to deflationism. The truth attribution that occurs in the first step makes no more assumptions about truth other than that it conforms to the deflationist schema DS. So there is no reason why the deflationist cannot accept this bit of reasoning as legitimate.

[87] To clarify, Quine makes two key moves in his solution to the paradox of non-being. First, he treats it as a semantic problem, rather than an ontological problem—that is, he treats it as solvable through semantic ascent and paraphrase. Secondly, he makes the specific proposal that it is solvable by treating names as short-hand descriptions. For our purposes, it is the first of these points that exemplifies the utility of the truth predicate. The second point is of less concern. Even if some alternative semantic proposal can equally well dispel the paradox, the first point will still stand that a device of semantic ascent is required to solve the paradox by attending to semantic matters.

[88] This inference relies on a crude version of Occam’s razor which recommends inference to the truth of the simplest theory. One could easily take issue with this principle. But for our purposes, it doesn’t really matter what the norms of theory choice really are. There just has to be some such norms. In that case, the point stands that the truth predicate allows us to ascend from first-order questions about the world to questions about norm-theory compliance.

5.2 Inflationism I: Rattan's cognitive inflationary model

We have now identified *four* applications of the truth concept that are recognizable to the deflationist. They are (i) expressing pluralities of statements, (ii) ascending to neutralize ontological commitment, (iii) ascending to paraphrase or reconceptualize the question of inquiry, and (iv) ascending to bring to bear the holistic norms of theory choice on first-order inquiry. Each of these functions may be incited by inquiry into the states of the world, and yet, they allow us to bring in considerations about thought and language.

Deflationism thus gives the truth concept a fairly impressive list of jobs. Indeed, this is more than is ordinarily supposed. At this point, one might reasonably ask: what *can't* the truth concept do, according to deflationism?

Before I give my own answer, it is worth looking at a rival proposal from Gurpreet Rattan. In his (2016), Rattan develops an account of the cognitive value of the truth concept that he calls the 'cognitive inflationary model'. His account is meant to serve (among other things) as a polemic against deflationism. It is worth outlining it here because it provides context for my view and serves as an illustrative contrast.

Until now, I have been speaking broadly of truth attributions to both linguistic and mental truth bearers, but Rattan is more narrowly focussed. His account pertains specifically to truth attributions to thoughts whose propositional content is explicitly formulated (2016: 231). For him, the paradigmatic truth attribution is of the form *that P is true* (e.g. *that water is H_20 is true; that factory farming is wrong is true*). Rattan argues that, on an intuitive understanding, these truth attributions have peculiar features that are not shared by other kinds. In particular, he argues that entertaining them "involves thinking with or entertaining the thought to which the explicit truth attribution attributes truth" (233). Thus, when I think *that factory farming is wrong is true*, I am simultaneously *employing* the thought *that factory farming is wrong* while also thinking *about* it. This mixed case of use and mention singles out explicit propositional truth attributions as special.

Rattan also has a dialectical motive for focussing on explicit truth attributions in thought. According to him, the deflationist position ascribes no cognitive value to the concept of truth when it is used in such thoughts; the concept only has cognitive value when it is used in inexplicit truth attributions (234, 244).[89] Hence, in Rattan's view, it is possible to challenge

[89] Besides deflationism, Rattan also has another opponent: the inflationary view of Collins (2007). In Collins's view, the concept of truth has cognitive value in inexplicit truth attributions because it allows us to represent the truth of truth-bearers whose contents we are unable to explicitly formulate. Rattan's cognitive inflationary model challenges the claim that this is the *sole* cognitive value of the concept of truth. Regardless, it still remains an open possibility that this is *one* (among many) roles for the concept of truth that is inextricably inflationary. See Moore (2020) for the case that this role cannot be performed by a deflationary concept of truth. However, I will not pursue this avenue here.

deflationism by simply locating some cognitive value in explicit truth attributions. Now, I have already given some reason to doubt that this strategy for undermining deflationism will work in general. Nonetheless, it is still instructive to look at the details of Rattan's theory in case the specifics are incompatible with deflationism.

So then, what is the distinctly inflationary role for explicit propositional truth attributions? According to Rattan, the cognitive value of explicit truth attributions is that they facilitate higher-order reflection on our own conceptual resources (228). The concept of truth allows us to turn our reflective gaze inward, contemplate our concepts and thoughts, and analyze the conditions under which they *apply* and are *true*. We can thus employ it to formulate and interrogate semantic hypotheses of the form *<P> is true if and only if Q* and *<F> applies to x if and only if x is G*. In short, the concept of truth allows us to perform *conceptual analysis*—specifically, *truth-conditional* analysis (234).

What's more, for Rattan, this activity of conceptual analysis is not merely aimed at discovering truths *about* our concepts and thoughts. On his picture, confirming a semantic hypothesis involves a "rational back and forth between intuitive judgments about examples and explicit (maybe partial) analysis of concepts"; we *employ* our first-order concepts in thinking about the world while simultaneously *mentioning* them to ascribe them semantic features (234–5). Conceptual analysis thus involves both thinking *of* and *with* the target thought or concept that is subject to analysis. (This is why it takes an explicit truth attribution; it exploits the fact that explicit truth attributions both use and mention the relevant thought or concept (238).) Therefore, not only do we gain semantic knowledge through conceptual analysis, but we also improve the clarity and mastery we have over our concepts as they figure into our first-order thoughts (234). By engaging in this higher-order reflection, we can thereby achieve a greater quality of justification for our first-order beliefs (234–6).[90] It is for this reason that the concept of truth is

[90] He writes, "The question is: how can reflection on concepts generate knowledge of the world? The short answer is as follows. Sometimes acquiring or improving knowledge requires clarity in understanding, clarity about what it is that one wants to know or know better. This clarity in understanding is provided by conceptual analysis, including some uses of truth conditional semantics, for the concepts and thoughts involved" (234).

consequential for our knowledge of the world.[91]

It is worth seeing some examples to see how this works in practice. Of course, not every first-order question demands higher-order reflection on concepts. Most inquiries have relatively well-understood methods for resolution, even when they are challenging to implement. But according to Rattan, there are some especially tricky cases where higher-order reflection is fruitful. He illustrates this point using examples from Williamson (2007) and Kennedy and Stanley (2009).

The first example concerns vague concepts in application to borderline cases. In his (2007), Williamson invites us to imagine that the planet Mars was once covered in water. But then, over time, the water molecules slowly left the atmosphere until Mars was totally desiccated. In this scenario, there will be a time when Mars is not clearly dry, nor clearly not dry—it will be a borderline case of dryness. Given this set-up, Williamson invites us to consider this question: *has it always been the case that Mars is either dry or not dry?*[92]

For Rattan, there are two features of this question that reveal the cognitive value of the truth concept. First, as Williamson goes to great lengths to argue, the question, as it is originally formulated, *is an object-level question about Mars* (2007: 23–31). It specifically asks of Mars whether it has a certain property—namely, the property of having always been either dry or not dry. By contrast, the question should not be construed as ultimately about thought, language, or concepts.

Nonetheless, even though this question is (arguably) not *about* thought or language, its second crucial feature is that it can only be properly answered if we attend to the semantics. We cannot resolve it through first-order empirical methods alone (such as counting the water molecules present at each time) without begging the question. This is because the rival resolutions to the question each employ different semantic theories for the logical operators and vague predicates involved. If, for instance, the semantic facts validate classic logic, then it will

[91] In his (2010), Rattan offers an alternative account of the cognitive value of the concept of truth. Instead of focusing on conceptual analysis, he focuses on the role of truth in 'critical reflective thinking'. (He also clarifies in (2016: fn.17) that he thinks of conceptual analysis as a species of critical reflective thinking.) The basic idea is that the concept of truth is valuable because it allows us to scrutinize our first-order beliefs in light of our norms of reasoning, while further scrutinizing those norms themselves in light of a standard of truth (2010: 12).

Like his (2016), Rattan's (2010) is also aimed to identify a role for the truth concept that is distinctly non-deflationary. But I share similar worries towards this account as I do to the 2016 account: it seems to me possible for the deflationist to acknowledge the role while keeping truth deflated. We have already observed that the deflationist can allow for the truth concept to serve as a means of semantic ascent for the purpose of checking theory-norm compliance. Can the deflationist also make sense of evaluating our rules and norms against a norm of truth? I don't see why not. A rule conforms to the truth norm if it delivers a lot of truths and not many falsehoods—that is to say, if it is reliable. Moreover, a rule is reliable if (by and large) it recommends the belief that it is raining (given the evidence) if, and only if, it is raining; it recommends the belief that it is sunny (given the evidence) if, and only if, it is sunny; and so on. In order to express the reliability of a rule in full generality, without running through each possible belief, we will need to mention truth. But, in doing so, we will only be using the truth concept to avoid an infinite conjunction. In other words, we'd be using the truth concept as a device of generalization, and nothing more.

[92] In semi-formal terms, the proposition at issue is *(∀ times t)((Mars is dry at t) or not-(Mars is dry at t))*.

be *true* that Mars has always been dry or not dry; but if they validate multi-valued or fuzzy logic, it will be indeterminate. Either way, the crucial point is that we cannot fully justify an answer to this question until we reflect on the semantic facts of the case. To this end, we must semantically ascend (using the concept of truth) and theorize about the truth-conditional contributions of the relevant concepts. If, by doing so, we can justify a particular semantic theory, then we can improve our justification of a given answer to the original question about Mars. In this way, we see how higher-order semantic theorizing can inform our views on questions that are ultimately about the world.

The second example exhibits a similar pattern, except that it concerns statements involving 'average' and their ontological commitments. Consider the statement 'the average American has 2.3 children'. Presumably, this could be true. However, if assenting to its truth incurs ontological commitment, then a surface-level reading of its structure suggests that there is an entity called *the average American* and that it instantiates the impossible property of *having 2.3 children.* But such commitments are clearly unacceptable by the lights of any realistic ontological theory.[93] Hence, there's a puzzle: how can we be justified to assent to such statements?

Kennedy and Stanley offer a solution in their (2009). They develop a semantic theory for phrases of the form 'the average *NP*' that avoids interpreting them as singular terms. Their theory employs the usual modelling assumptions of formal semantics in the Fregean tradition: it explains semantic composition via functional application, and it takes objects, functions, and truth values as the basic ingredients for semantic values. For Kennedy and Stanley, the trick is ultimately to treat 'the average' as its own lexical item that denotes a function with the following operation: given a class *C* (e.g. Americans) and a measure function *f* (e.g. having *n* children), it returns a truth value upon input of a number *n* (e.g. 2.3); it returns true just in case *n* equals the sum of *f(c)* for all *c* in *C* divided by the cardinality of *C* (2009: 614). Apart from this, the remainder of their theory involves conforming this idea to the constraints that are independently imposed by syntax and compositionality.

Details aside, the interest in Kennedy and Stanley's account for Rattan and us is the distinctive way that the concept of truth is employed to serve broader theoretical aims. One kind of question that we may ask about the world involves *averages*—e.g. *whether the average American has 2.3 children.* Now, according to Rattan, "what makes these questions special is the need to ascend to the meta-level and begin to think about just what thought it is that one is thinking and what exactly would constitute justification for it—an analysis or semantics for the concept *average* is required" (2016: 236–7). But when we think about the *semantics* of these thoughts, our broader aim is still to answer the first-order question about real-world averages.

[93] There are also independent semantic reasons for denying that '*the average American*' functions as a singular term, even if we construe its alleged denotation as a fictional entity. See Kennedy & Stanley (2009: 596–8).

And thus we see how the concept of truth is imperative. It allows us to ascend for the sake of truth-conditional analysis while keeping our eye on the world.

I said earlier that my own inflationary role for truth contrasts with Rattan's. But before I make any criticisms, let's state upfront what Rattan's model gets right. It is certainly true that the concept of truth—and its kin, *<refers>* and *<applies>*—are indispensable to representing the truth-conditional semantic facts. They therefore have a distinctive role to play in our thinking about semantics: they allow us to conceptualize and interrogate the truth conditions of our first-order thoughts. Moreover, it is plausible that higher-order reflection on semantics has *some* role in our broader efforts to learn about the world, and not just the narrow part of the world that concerns human thought and language. It is thus sensible to focus on this reflection to understand the functions of the truth concept. Nonetheless, I have two worries about how Rattan's account implements this idea.

My first worry stems from a wider theoretical disagreement that concerns the underpinnings of Rattan's picture. Rattan's account assumes that it is possible to gain non-trivial insight into the nature of things by interrogating the semantic contents of the concepts (and words) that we use to think (and talk) about those things. That is, it assumes that conceptual analysis can advance the aims of first-order inquiry. But this assumption is itself highly contentious. Indeed, like many post-linguistic/conceptual turn philosophers, I am skeptical of the idea that we can ever learn much about extra-mental and extra-linguistic reality through direct semantic reflection. The picture that I prefer instead would assign a relatively modest role to semantic analysis in first-order inquiry. However, it would take me too far afield to make this case here.[94] So at least within the confines of this chapter, I will have to leave this objection inconclusive.

My second, more considerable worry concerns the account's dialectical effectiveness against deflationism. As long as we focus on semantic analysis, we remain in territory that deflationism is poised to encroach. Rattan locates the cognitive value of the truth concept in its applications in semantic analysis: the truth concept and its kin allow us to entertain and justify *non-disquotational* semantic hypothesis of the forms *<P> is true if and only if Q, <a> refers to b,* and *<F> applies to x if and only if x is G.* He is surely right that the concept of truth does this. However, I'm not so sure that the *deflationist* conception of truth is inevitably opposed to this application. This raises the question: can the deflationist recognize a role for truth-conditional analysis?

In fact, we have already seen some reason to think that they can.[95] (And the fact that truth-conditional analysis takes *explicit* truth attributions is insufficient to prove otherwise.)

[94] See Taylor (2019).

[95] It is noteworthy that there are philosophers who are deflationists about truth and yet they advocate for linguistic and conceptual analysis to serve broadly metaphysical ends. Thomasson (2014b) is a case in point.

Recall, from earlier, that one deflationist-friendly use of the truth concept is semantic ascent for paraphrasing or reconceptualization. By availing ourselves of this application, we see that the following line of reasoning is entirely permissible by deflationist lights. First, we begin with a first-order thought *that P* (let this be the first-order question of ultimate concern). We then employ the deflationary truth concept to attain the biconditional *P if and only if <P> is true.* Once we've ascended and turned our reflective gaze towards <P>, we can consider appropriate *paraphrases, translations,* or *alternative ways of conceptualizing* <P>. Now suppose that we arrive at a suitable equivalent representation <Q>; we then get the equivalence *<P> is true if and only if <Q> is true.* Finally, with a second use of the deflationist truth concept, we can semantically descend to arrive at a non-disquotational analysis of <P>'s truth conditions, *<P> is true if and only if Q.*

The point is, deflationists about truth have a way of simulating the process and results of truth-conditional analysis.[96] Under a deflationist conception, this process will be understood as a matter of *paraphrasing* the analysandum in a more perspicuous way and then semantically descending.[97] It is true that deflationists understand the *grounds* for this process differently than do inflationists. (For deflationists, justifying a semantic hypothesis is ultimately a matter of justifying a paraphrase as appropriate or apt. To this end, they may appeal to shared conceptual roles, shared causal history, or other shared features of use. Whereas inflationists see semantic analysis as concerned directly with the relations between representations and real-world objects. See chapter two.) But if we put the metasemantic differences aside (for the moment), we see how the two can look quite similar in practice.

To see this, consider Kennedy and Stanley's account of 'average' again. As previously stressed, their semantic analysis of 'the average *NP*' has the wider theoretical purpose of showing how statements like 'the average American has 2.3 children' can be true given a realistic ontological outlook, despite its apparent reference to *the average American*. In this respect, their semantic theory is intended to serve exactly the same purpose for metaphysical inquiry as Russell's theory of descriptions does for Quine (1948) regarding 'Pegasus'. To reiterate the familiar point, Quine analyzes 'Pegasus does not exist' to be true just in case there is no winged

[96] Indeed, there is a sizable body of work defending the idea that deflationism about truth is compatible with truth-conditional theories of meaning; see Williams (1999), Horisk (2008), and Burgess (2011).

[97] One might object that this understanding of truth-conditional analysis is *ad hoc*. Deflationism aside, we shouldn't understand truth-conditional semantics as fundamentally about translation; it is first-and-foremost about real symbol-object relations and it shouldn't be reconstrued otherwise. Given Rattan's remarks on (2016), p. 234–5, I'd expect this to be his response. Speaking for myself, I agree with this objection; the crux of chapter two is that it is *ad hoc* and backwards to treat interlinguistic translation as more fundamental to semantics than reference. However, I do not know how to *prove* this point to the satisfaction of the deflationist. And my present point is not that deflationism's understanding of truth-conditional analysis is *correct.* Rather, it is that it is *viable*, and hence it is possible for deflationism to reap the rewards of the cognitive value that Rattan ascribes to the truth concept. If that's right, then the dialectical upshot is that Rattan's polemic against deflationism doesn't succeed. In order to challenge deflationism, we need to find a role for the truth concept that is even more inflationary. And that's what I intend to do in the next section.

horse caught by Bellerophon. This ascription of non-disquotational truth conditions is the result of something like semantic analysis.[98] But for Quine, it is also nothing more than finding an adequate paraphrase for ‘Pegasus does not exist’, and the paraphrase is effective precisely because it dispels the apparent reference to Pegasus, thus displaying the statement’s ontological neutrality. And as we have noted before, semantic analysis done with this aim and with these presuppositions is entirely compatible with deflationism.

Now, Kennedy and Stanley’s account is more complicated because they are working under additional theoretical constraints. Besides its role in discerning ontological presuppositions, semantic analysis, for them, is also supposed to represent the compositionality of natural language in accordance with its actual syntax. Given these constraints and the usual modelling assumptions of formal semantics, their account singles out ‘the average’ as semantically simple and assigns it a function as its denotation—from classes, measure functions and numbers to truth values. The immediate *ontological* upshot of this theory is that it no longer treats ‘the average American’ as a singular term, thereby ridding ‘the average American’ of the illusion of ontological commitment.[99] But it also means that the interpretation of the ‘the average American has 2.3 children’ will include exotic mathematical objects as assigned semantic values (case in point: the denotation assigned to ‘the average’). And given how far afield the formal capture is from the statement’s disquotational truth conditions, one might wonder whether it is really compatible with a deflationist’s understanding of truth.

Nevertheless, I would like to suggest that it is. Like Quine’s ‘Pegasus’, the deflationist would regard Kennedy and Stanley’s exercise in formal semantics as a matter of paraphrasing their target sentence into another idiom, abiding by the various constraints imposed by their various explanatory ends. In this case, their aim to model the compositionality of natural language commits them to couch their interpretation in a mathematical idiom, including the devices of function and lambda abstraction. They must also ensure that their interpretation respects the conceptual roles of the original statement’s significant parts. But for the deflationist, this is all just a matter of making sure that the paraphrase is the right one for the task at hand. Kennedy and Stanley’s interpretation is chosen by how well it represents natural language compositionality and reveals ontological commitment. But most importantly, this explanatory endeavour does not also aim to speak on the metaphysical characters of the fundamental semantic properties, namely truth and reference. As such, it is neutral towards the more fundamental division between deflationism and inflationism.[100]

[98] Given his rejection of the analytic/synthetic distinction, the historical Quine would protest against this characterization. But let’s set that aside, since deflationism about truth and the wholesale rejection of semantic analysis are separable positions.

[99] But not entirely. Since, according to their account, ‘the average’ denotes a function that takes *numbers* as arguments, their theory vindicates an ontology that includes numbers (641–2).

[100] For more on this, see Burgess (2011) and K. Taylor (2019: ch. 2).

So, then, where does this leave the dialectic regarding Rattan's account? As I have already stated, I think that Rattan's account is broadly right about the function that it ascribes to the truth concept: the truth concept is a means for higher-order reflection on the semantics (specifically truth-conditional content) of our first-order thoughts. Moreover, it is reasonable to expect that this function has some part in serving the aims of first-order inquiry. (However, I'm personally doubtful that first-order questions can be straightforwardly answered through conceptual analysis.) So on these points, Rattan's account is promising. But with that said, I am skeptical of the claim that Rattan has identified a distinctly *inflationary* function of the truth concept. For all that is said, a deflationary truth concept can capture the cognitive value espoused by Rattan. So as a polemic against deflationism, Rattan's account falls short. If we want to prove that the debate between inflationists and deflationists is consequential to the concerns of first-order inquiry, we will need to look elsewhere for a genuinely *inflationary* function of the concept of truth.

5.3 Inflationism II: representing semantic determinants

Above all else, the last section reveals how incredibly *flexible* the deflationary truth concept is. Indeed, deflationists often make a point of imitating the inflationist's claims regarding most points of contention. Hence, they offer deflationist theories of vagueness, semantic indeterminacy, and non-factualist discourse (Field 1994b; Horwich 1998a; Leeds 2000), deflationist theories of meaning and truth-conditional semantics (Horwich 1998b; Field 1994a; Williams 1999), deflationist theories of the explanatory role of truth concerning successful action and empirical prediction (Leeds 1995; Horwich 1998a; Field 2005; Maddy 2007), and so on. Their overall effort is to minimize the distance between the two views, thereby arguing that we can say all that we want with a less pretentious notion of truth.

However, if deflationism is to remain a distinctive view, it must really offer a less metaphysically-loaded notion of truth. Hence, there is still one point where the deflationist and inflationist must distinguish themselves: they must each offer different accounts of truth's nature.

As we have seen, the deflationist's metaphysical thesis is that the account of truth's nature must ultimately remain trivial. This means that, in the final analysis, there's nothing deeper to say about truth than what is provided by the total of instances of DS (that is, notwithstanding the supplemental claims that one can make through paraphrase or reconceptualization). Thus, according to the deflationist, it's a mistake to search for any deeper explanation of why *'Socrates is wise' is true if and only if Socrates is wise.* This fact is supposedly explicable by the logical workings of the truth concept (or predicate), which, in turn, is explained by its role as a device for semantic ascent and descent. This truth-conditional fact is not to be explained by any further relations between, for instance, the subject term ‘Socrates’ and Socrates himself.

To echo §2.3, this negative metaphysical thesis entails further deflationary theses concerning the semantic relations that pertain to subsentential expressions (or sub-propositional concepts).[101] Like truth, the deflationist must also say that *reference* is (in some sense) ‘insubstantial’, and similar claims must be made about the semantic relations for other subsentential expressions. In particular, they must say that the facts of reference (within one’s home language) are entirely captured by the reference disquotational schema,

(RDS) ‘*a*’ refers to *a* (if *a* exists).

(Again, to deal with context-sensitive and foreign expressions, the deflationist will also appeal to translation and paraphrase. But we can mostly ignore these technicalities here.)

The significance of the deflationary theory of reference resides primarily in what it denies. Since it claims that RDS instances are entirely grounded in the logic of ‘refers’, deflationism entails that there cannot be any *further* explanation for these facts of reference. In particular, there cannot be any deeper explanation as to *why* ‘Socrates’ refers to Socrates, or what this reference relation consists of.

Deflationism thus stands in opposition to two kinds of projects encompassed by traditional theorizing about reference. First, they are committed to rejecting all purported analyses (or a posteriori reductive accounts) of the reference relation that attempt to spell out its nature in more fundamental terms (Horwich 1998b: 123; Field 1994a: 260; Leeds 1995: 15). They are thus committed to denying the descriptivist theory (e.g. Frege 1948; Russell 1919), the causal theory (Stampe 1977), the informational theories (Dretske 1981; Fodor 1987) and the teleosemantic theories (Millikan 1984; Neander 2017). For the deflationist, reference is not the kind of relation that has an underlying nature that is amenable to such explication.

Besides this, deflationism is also committed to rejecting the more modest projects of inflationary metasemantics, which earlier I called the ‘selective explanations of reference’. Generally speaking, the *inflationary* approach holds that there are non-trivial answers to the determination questions of truth-conditional metasemantics—that is, questions of the form: *why is it that ‘a’ refers to b (as opposed to anything else)?* For inflationists, these questions find their answers in the more fundamental relations that subjects bear to objects concerning how they employ their terms and their relations to their environment. (Perhaps these relations are broadly causal, or perhaps they involve descriptions grasped by subjects.)

[101] My claim that truth deflationism entails reference deflationism is uncontroversial in the literature. For example, it is explicitly endorsed by Field (1994a: 261) and Horwich (2005: 74). The basic reasoning behind the entailment is that *if* it were possible to give a non-deflationary account of reference, then this account could be transcribed into a non-deflationary account of truth. This argument is spelled out by D. Taylor (2019).

But, for the deflationist, there is no deep explanation as to why particular reference relations obtain. Deflationists are generally opposed to all non-trivial answers to the determination questions of metasemantics (that is when metasemantics targets the *truth-conditional* semantic facts; from now on, whenever I speak of metasemantics, I am specifically referring to its *truth-conditional* variety).[102] Hence, for them, there cannot be any explanation as to why, e.g., 'water' refers to water (as opposed to, say, twin-water) besides citing the disquotational features of 'refers'.[103] It would be a mistake to explain this by citing our causal interactions with water (and lack of interaction with twin-water). Our causal interactions may explain a variety of things, but, according to the deflationist, they cannot explain *reference fixation.*[104]

Given these negative claims from the deflationist, we are now finally in a position to split the difference between deflationism and inflationism concerning the roles of truth. In sum, inflationism permits for there to be substantive explanations of reference and truth conditions, whereas deflationism precludes them. Therefore, one thing that the truth concept—and its kindred concepts, <refers>, <applies>—*cannot do* under a deflationary conception, but *can do* under an inflationary conception, is be employed for the purpose of substantive metasemantic theorizing. For the deflationist, there ought to be no role for semantic ascent to reflect on *why* sentences or thoughts have the truth conditions that they have, or *why* words or concepts have the referents that they do. For them, these 'why'-questions must receive trivial answers. Whereas, for the inflationist, the concept of truth (and concept of reference) may facilitate non-trivial answers to the various questions of truth-conditional metasemantics. It follows, then, that a distinctly inflationary role for the concept of truth is *semantic ascent for metasemantic purposes.*

Let's call this inflationary function for the truth concept 'representing semantic determinants'. Before I attempt to detail how it might work in practice, let's describe it in the abstract.

Ultimately, the plan is to articulate an inflationary role for the truth concept that is instrumental to first-order inquiry. So let's suppose that we are inquiring into the nature of some worldly entity X (e.g. water or wrongness). And let's further suppose that we have explicitly formulated the hypothesis under investigation that *X is F*. The concept of truth will have an inflationary role to play if truth-conditional metasemantic reflection has any part in our investigation as to whether *X is F.*

[102] Deflationists need not be opposed to substantive metasemantics when the semantic facts are construed as something other than truth-conditional. Indeed, deflationalists typically trade in truth-conditional metasemantics for some alternative conception of the foundations of meaning, such as a 'use' theory of meaning (Horwich 1998b).

[103] This assumes that the mentioned word is suitable for disquotation. If it isn't, then (as argued in chapter two) the causal connections between word and object may be cited as a relevant consideration for choosing an appropriate *paraphrase.* But they do not explain reference *directly*.

[104] See Field (1994a) section 4; Taylor (2017).

To engage this role, we first semantically ascend to formulate the higher-order question of whether *'X is F' is true* and note that *'X is F' is true* if and only if '*X*'s referent satisfies '*F*'. This gives us a view of the semantic facts, that *'X' refers to X* and *'F' applies to the Fs.* Once the semantic facts are in view, we are then able to consider the underlying mechanisms that *determine* or *ground* the semantic facts. We can entertain and justify metasemantic hypotheses of the form *'X' refers to X in virtue of 'X' bearing R to X* or *'F' applies to the Fs in virtue of 'F' bearing R* to the Fs.* We thereby use the concepts of truth, reference, and application to represent the *determinants* of the truth-conditional semantic facts.

Finally, suppose, for the sake of argument, that these metasemantic theories have some part to play in our original inquiry. That is, suppose that our understanding of the determinants of reference for '*X*' or '*F*' plays some crucial dialectical role in justifying an answer as to whether *X is F*. (How this might work in practice will be considered in the next sections.) In that case, we will have found a function for the truth concept that is indisputably inflationary.

Notice how this function for the truth concept differs from Rattan's proposal. According to Rattan, the truth concept is used to facilitate theorizing about the *semantics* of our first-order thoughts; whereas, according to my proposal, the truth concept may also be used to enable theorizing about the *metasemantics* of our first-order thoughts and sentences. This difference is important for a variety of dialectical reasons. For one, my proposal may be more appealing to those who subscribe to the theory of direct reference, since they are generally doubtful that semantic analysis can afford insight into the nature of referents. My proposal is also better positioned to distance itself from deflationism since this potential function cannot be achieved under a deflationary conception of truth. Since the deflationist is firmly committed to trivializing truth-conditional metasemantics, they cannot recapitulate this function without forfeiting their view.

So far, I have argued that this potential function represents a wedge issue between deflationism and inflationism. We have found that *if* truth-conditional metasemantic reflection ever plays a part in informing or justifying our views on the world, then *ipso facto* we are presupposing an inflationary conception of truth. This is enough to establish that the inflationism-deflationism divide could *in theory* make a difference in how we inquire about the world. So if my aim were only to argue that there's a *potential* connection between the different conceptions of truth and first-order inquiry, then I would have now accomplished it. However, this doesn't yet show that this division matters in practice. After all, the deflationists will probably insist that the function I've identified is idle. They will argue that rational inquiry can, and should, proceed without the need for higher-order reflection on semantic determinants.

Given this dialectical situation, we see that the debate over the functions of the truth concept now turns to another question: is there ever any occasion in which truth-conditional metasemantic reflection contributes to our beliefs about the world? Suppose, for example, that we are investigating whether *water is essentially H_20* or whether *factory farming is wrong.* To answer these first-order questions, is there any value in semantically ascending to reflect on the metasemantics of the key terms? Is any epistemic advantage gained from considering the grounds of reference for 'water' and 'wrongness'? In short, does reflecting on the *mechanisms* of reference give us any insight into the nature of *referents*?

In what remains, I would like to raise the stakes by motivating the inflationist's side on this question. I would like to give credence to the idea that metasemantic reflection can aid first-order inquiry. It seems to me that there are episodes in the history of philosophy that can plausibly be interpreted as exhibiting this pattern. I will briefly outline two of them.

But before I do, there are a few comments on the dialectical situation that are worth keeping in mind. First, I do not intend to *prove* the inflationist's position through this roundabout method of giving inflationary glosses on first-order philosophical disputes. Doing so would require a much more substantial discussion of the applications than I can afford to offer here. In effect, I would need to enter into the first-order disputes themselves and defend a particular position—a task far beyond the reaches of a chapter on the concept of truth. Rather, my purpose is to illustrate how the debates about truth can make a difference in how we investigate the world. The best way to do this is by motivating the applications that are distinctly inflationary.

Secondly, it is also worth emphasizing that the question is highly delicate. It is entangled with several larger issues in the philosophy of language and metaphilosophy, so there is little that I can say that won't be contentious. But it is also not my intention to defend these applications to the ends of the earth. Instead, the best that I can do is be forthright about my assumptions and remind the reader that if they don't share them, there may be plenty of other applications that are more to their liking.

Thirdly, the range of epistemic roles for metasemantic reflection will no doubt depend on whether the mechanisms of semantic determination are broadly internal or external. For according to internalism, the metasemantic facts are open to *a priori* reflection and readily available for direct application in thinking about objects. Under this picture, it might be thought possible to justifiably *infer* first-order claims by consulting the grounds that determine their content (*viz.,* through conceptual analysis). However, if content is determined along externalist lines—as I think it is—then the applications of metasemantic reflection cannot be so ambitious. Since I'm sympathetic to externalism, the applications that I exhibit will be relatively indirect. I will not claim that the cognitive role of metasemantic reflection is to provide a source of *evidence* from which one can *infer* the truth of any first-order claim. (When a metasemantic theory plays a role in justifying a first-order claim, it will do so in conjunction with other sufficient first-order evidence.) Nor will I claim that the role played by metasemantic reflection is *indispensable* to the justification of our first order beliefs. Indeed, first-order beliefs require first-order evidence. All that I claim is that metasemantic theories can (on occasion, in philosophical contexts) buttress the justification of our first-order beliefs—by dispelling certain puzzles, defending against certain objections, and providing broader explanations of how our theorizing works. Nevertheless, this limited role would still demonstrate that metasemantic reflection can make real epistemic differences. That's enough to show that the accompanying concepts have cognitive value—that the concepts of truth and reference have *inflationary* cognitive value.

Finally, I do not mean to suggest that the overall case for inflationism rests solely on the applications of the truth concept in *first-order* inquiry. For inflationism to be true, it's enough that the facts of reference and truth conditions be grounded in real-world relations between subjects and objects. This may be argued in a variety of ways. (It is most commonly argued by showcasing a metasemantic theory's explanatory pay-off for cognitive science.) The way explored here is just one route to inflationism. I explore it because it has not yet received as much attention as the other well-worn inflationist arguments, and because the overall interest of this chapter is the connection between truth and first-order inquiry.

5.4 Scientific progress and essentialism

The first episode concerns the causal theory of reference and its history. It is well known that the causal theory, developed by Kripke and Putnam, was originally instrumental in defending a number of realist theses, specifically concerning the subject matter of science and the nature of necessity. The causal theory is also widely considered a paradigmatic example of inflationist metasemantics. It is thus reasonable to expect that the theory's historical applications will exhibit the desired inflationist pattern.

Following Putnam (1973, 1975), we can briefly characterize the causal theory as comprising the following claims about the metasemantics of substance and natural kind terms. First, it holds that reference for these terms is *not* determined by the descriptions, conceptions, or theories held by the subjects who wield them (Putnam 1973: 700–4, 1975: 143–4; Salmon 2005: 94). Instead, it is ultimately determined by the samples and instances that subjects actually perceive and otherwise causally interact with. Putnam explains this in terms of ostensive definitions that demonstratively refer to actually perceived samples (1973: 707, 1975: 148–9). Secondly, the semantic intension (i.e. function from worlds to extensions) of a substance or natural kind term is thereafter determined by the *objective similarities* that things bear to the perceived samples or instances. Again, intensions are *not* determined by the internal states of subjects. In the case of water, Putnam argues that the reference-fixing mechanism involves whatever underlying physical relation that determines two sample liquids to belong to the same kind (1973: 702–3, 1975: 142). He specifically represents it with the formula:

> *For every world w, for every x in w, x is water if and only if x bears same$_L$* [the relation that determines sameness of kind for liquids] *to the entity referred to as 'this'* [a sample perceived and demonstrated] *in the actual world* (1973: 707, 1975: 149).[105]

We know, as a matter of empirical discovery, that this relation consists of shared chemical composition.

I'm bypassing a tremendous amount of detail (see Salmon 2005: ch. 4). But our interest lies in the applications, not in the theory itself.

Perhaps the most famous application stems from the theory's historical ties to scientific realism. Roughly, the theory gives substance to the realist idea that the entities of science possess their natures independently of our theorizing.[106] Indeed, Putnam's metasemantics *presupposes* the realist thesis that the samples of a natural kind or substance bear objective similarities amongst each other. But on top of this presupposition, the theory also *explains* how we can secure reference to substances and kinds without thereby delineating the substances and kinds by our conceptions of them. And within its historical milieu, this explanation of reference played a pivotal role in defending the realist conception of science against its anti-realist opponents (e.g. Kuhn 1962 and Feyerabend 1962).

This application can be made more concrete by considering the problem of inter-theoretic reference and truth (Putnam 1973: 153–7). A central pillar of scientific realism is that science *progresses.* This means (among other things) that modern stages of scientific theorizing offer an improved understanding of (some of) the same substances, kinds, and entities that were discussed in earlier stages—e.g. that modern chemical theory provides a better theory of water than the folk theory of ancient Greece. For the realist, scientific progress does not merely consist of making more accurate and plentiful predictions; it also involves improved knowledge *of* the world. But in order for modern scientists to have improved upon their predecessors' understanding of, e.g., water, it must be that their respective 'water' theories *share the same*

[105] This is a stylistic variant of the original text. In particular, I've retained Putnam's convention of representing the semantic intension of 'water' by *using* the term 'water' on the left-hand side of the biconditional. To make the meta-representational nature of the claim more explicit, we write, "*for every world w, for every x in w, 'water' (as actually used) refers to x with respect to w if and only if…*".

A more substantial variation is provided by Salmon (2005): "It is necessary that: something is a sample of water if and only if it is a sample of the same *actual* substance that *this* is *actually* a sample of" (2005: 145). The advantage of Salmon's formulation is that, unlike Putnam's, it does not quantify over world-bound slices of objects (i.e. *x-in-w*). This is philosophically important because world-bound slices of objects arguably have all of their properties essentially, whereas ordinary objects (which are quantified over in Salmon's formulation) do not. So unlike Putnam's formulation, Salmon's does not risk tempting us into a cheap, but ultimately irrelevant, kind of essentialism.

[106] Scientific realism has many dimensions, but for our purposes, we can simplify by focussing on the claim that the entities of science (objects, properties, kinds, substance, etc.) possess their natures independently of our theorizing.

subject matter. And yet, the pre-modern conception of water is vastly different from the modern one. There may be little overlap between the respective *theories* or *descriptions* of water. So how is the sameness of subject matter secured?

From the anti-realist's perspective, this puzzle looks to be a formidable challenge to the realist's point of view (1973: 153–4). However, Putnam's theory has the virtue of defusing this challenge by *explaining* how inter-theoretic reference is possible. For according to his account, we mustn't look to the theories or descriptions produced by individual scientists to ground the reference of their terms. Instead, we should appeal to the fact that each scientific generation is causally interacting with the same natural environment. Given Putnam's metasemantics for substance terms, the reference for 'water', as used by a community of scientists, is determined by the samples they interact with as they develop their theory and the chemical composition of those samples. Hence, as long as each scientific generation interacts with samples of the same chemical kind, their respective theories will share the same subject matter. In this way, Putnam's theory explains how successive theories can improve our knowledge of a fixed stock of substances and kinds, whose natures are theory-independent.

Putnam's inflationary metasemantics thus makes a crucial contribution to the defence of the realist view of scientific entities. It does this by explaining how an otherwise-baffling feature of realism is possible. By demystifying inter-theoretic reference and progress, Putnam's metasemantics thereby allows for realism to enjoy a greater degree of rational entitlement than it would have otherwise. It is for this reason that I say that inflationary metasemantics plays a part in justifying the view.[107] As it happens, a realist view of the grounds of reference reinforces a realist view of scientific referents.

The deflationist, by contrast, cannot recognize this application of the concepts of truth and reference. Suppose that someone were to attempt to defend realism from the deflationary perspective, and so they needed a story about scientific progress. As per their deflationism, they cannot explain inter-theoretic reference (and thereby progress) by appealing to the shared grounds of reference between successive scientific generations, since deflationism entails that there are no such grounds. Instead, they would have to say that past uses of the scientific vocabulary translate into the present vocabulary—e.g. ancient Greek 'water' ('ὕδωρ') translates

[107] To clarify, I am *not* claiming that one can justifiably *infer* scientific realism from a prior belief in Putnam's account of reference, or that Putnam's account of reference constitutes *evidence* for scientific realism. Arguably, it does not, since (as previously mentioned) Putnam's metasemantics *presupposes* a realist metaphysics of substances and kinds. I am claiming that it makes a different, indirect sort of impact to the epistemic credentials of scientific realism: namely, it serves to counter the would-be challenges to realism posed by Kuhn and Feyerabend (the problem of inter-theoretic truth and reference).

Moreover, I am also not claiming that an explanation of inter-theoretic truth and reference (and hence, progress) is sufficient to defend realism against all of its anti-realist competitors. After all, there are some anti-realist opponents that will uphold inter-theoretic reference—such as the view that past and present theories are referring to the substances and kinds delineated and constituted by an ideal hypothetical theory. I am only claiming that an explanation of inter-theoretic truth and reference is an important *part* of the overall justification of realism.

as our ‘water’. However, it’s difficult for the deflationist to justify these translations without being *ad hoc*. They cannot say that ancient Greek ‘water’ (‘ὕδωρ’) and our ‘water’ translate on the basis of shared reference (§2.3). Even worse, they cannot appeal to shared causal origin without losing their distance from inflationism.[108] And without these materials, their view has little else to offer to *explain* inter-theoretic reference. The deflationist may insist that there is such a thing as inter-theoretic reference and progress, but they cannot give a decent explanation of how it is possible. For this reason, their view cannot capture the cognitive value that’s achievable by inflationism.

If this is correct, then inflationary metasemantics can play a distinctive role in defending a high-level claim about the nature of scientific entities—namely, that they are theory-independent. This is enough to show that the concepts of truth and reference can have inflationary cognitive value. But it doesn’t yet show that they have inflationary cognitive value *for first-order inquiry*. One could still wonder whether realism ever plays a role in justifying any particular first-order claim—e.g. whether *water is essentially* H_20. To bring Putnam’s discussion into contact with our overarching topic, we need to find a connection between the realist’s conception of science and the epistemology of scientific practice.

On this question, I must be much briefer than the subject deserves. But I trust that it shouldn’t be too controversial to claim that there is such a connection. The scientific realist views the scientific enterprise as interrogating a world of mind-and-language-independent entities with essential natures that are not determined by our theorizing. Such a conception of the subject matter of science lends itself to certain methodologies, rules of inference, and epistemic practices that are not equally favoured by anti-realist conceptions. To mention a renowned case in point, the realist is prone to taking a much more favourable attitude towards inference to the best explanation than the anti-realist, especially when unobserved entities or hidden essences are involved (Boyd 1983; Psillos 1999: ch. 4; Lipton 2004: ch. 11).[109] A realistic attitude towards chemical science (for instance) may thus play a part in warranting an inference to an essential chemical structure for a given substance to explain its common observable traits.

Suppose it is right that scientific realism differentially sanctions particular scientific methods. In that case, the realist’s conception of science will also have a part to play in justifying the first-order beliefs that are warranted by those methods. (That is, there will be some first-order propositions that will enjoy more rational support when given a realist backdrop than they would when given an anti-realist backdrop. The propositions that are recommended by IBE are a case in point.) Now, since inflationary metasemantics can play a role in defending scientific realism *à la*

[108] To reiterate the main line of chapter two, it is possible in principle for the deflationist to ape the metasemantic explanations of the inflationist, but in doing so, they render their position redundant and *ad hoc*.

[109] According to the realist, IBE should be understood to warrant *belief* in the (approximate) truth of the favoured hypothesis—as opposed to some weaker intentional attitude (cf. van Fraassen 1989: ch. 6–7).

Putnam, we find that there's the desired metasemantic-epistemological link.[110] Through this example, we thus see how inflationary concepts of truth and reference can have an impact on the justification of our first-order scientific beliefs.

A similar lesson can be drawn from history's other famous application of the causal theory of reference, namely, the defence of essentialism. It is well known that causalist metasemantics played *some* key dialectical role in defending particular claims about essences, such as that *water is essentially* H_20—or better, that *necessarily water is* H_20. However, the precise nature of that role is controversial (c.f. Salmon 2005). In this brief vignette, I aim to stick to what (I hope) is relatively uncontroversial while highlighting the importance of inflationary metasemantics.

First off, it is crucial not to overstate the metaphysical aspirations of the causal theory of reference. As Salmon (2005) shows, we should not expect to *derive* non-trivial essentialist claims from the theory of reference alone. Instead, we ought to think of the metasemantic theories as playing a corroborative role in defending essentialism. They can aid us in establishing essentialist conclusions, but only when taken in conjunction with independent metaphysical and empirical premises.

Specifically, the causal theorist's argument for water being *necessarily* H_20 can be broken down into three steps (Salmon 2005: 166–7). It begins with an assertion of Putnam's metasemantic theory applied to the substance term 'water':

> (i) Necessarily, something is a sample of *water* if and only if it bears the $same_L$ relation to the sample referred to as '*this*' [a sample perceived and demonstrated] in the actual world.

Next, there's an empirical premise about the composition of the demonstrated sample:

> (ii) *This* (liquid sample) has the chemical structure H_20.

Finally, there must be a premise to the effect of:

> (iii) Being a sample of the same substance as something [i.e. bearing $same_L$] consists in having the same chemical structure.

[110] Once again, this must be understood as the more moderate claim that inflationary metasemantics helps to buttress realism against certain anti-realist challenges—not as the bolder claim that it provides positive evidence for realism.

From these three statements, we get the desired result that, by necessity, every sample of water has the chemical composition H_20.[111]

By formulating the argument in this way, we see how Putnam's metasemantic theory figures into the defence of the first-order essentialist conclusion. The theory acts as an intermediary between a particular empirical premise (ii), a general metaphysical premise about substances (iii), and the desired conclusion about water. Salmon argues at length that the conclusion about water's essential nature cannot be derived without the help of premise (iii), and premise (iii) is a non-trivial essentialist thesis about substances (2005: ch. 6). So the argument incorporates a contentious metaphysical assumption that is independent of the theory of reference. Nonetheless, even if we cannot establish the essentialist conclusion without the help of auxiliary essentialist premises, Putnam's inflationary metasemantics still has a place in the argument.

Granted, this argument is only one route to the conclusion that necessarily, water is H_20. One could still wonder whether this particular line of reasoning, with its detour through inflationary metasemantics, is inextricable to the defence of essentialism.

To echo my stage-setting from earlier, I do *not* mean to claim that we cannot establish the first-order result without the help of Putnam's account of reference. Clearly, we can. For instance, we can also reason directly about water and its properties, without semantic ascent. In that case, we would answer the necessitation question by reasoning that *water is necessarily H_20* because *to be water is to be H_20*, and then by providing a further, essentialist story about this form of explanation. This is the most foundational way to reach essentialist conclusions because first-order claims require first-order evidence. Moreover, since this explanation doesn't mention truth, it is metasemantically neutral and hence entirely acceptable to the deflationist.

Nevertheless, I still claim that the metasemantic detour can make an important contribution to the overall defence of essentialism. In my view, the value of metasemantic theory in this context is that it can help explain an otherwise baffling fact about the *epistemology* of essence attributions: that they can be *a posteriori.* Bear in mind that before Kripke and Putnam, the widespread, centuries-old dogma was that all necessary statements are knowable only *a priori*. So at the time that the causal theory arrived on the scene, a proper explanation of the *necessity* of water being H_20—a patently *a posteriori* fact—would have required an explanation for how this dogma is false. Putnam's causal theory provides this precisely because it explains

[111] This formulation is paraphrased from Salmon (2005: 166–7). Departures from the original text were done to preserve consistency with my exposition from earlier.

how reference fixation depends on features that are external to the thinking subject.[112] On his theory, the reference of ‘water’ depends, in part, on the actual chemical composition of the samples that subjects interact with, and this composition is entirely beyond their *a priori* ken. Contrary to the pre-Putnam-Kripke paradigm, reference is not fixed by internal facts that each subject grasps in virtue of their conceptual competence. For this reason, competence with the concept of water does not entail knowledge of the underlying structure that binds the samples together as a common kind. *A fortiori*, competence with the concept of water does not entail *a priori* knowledge of water’s essence.

If this is right, then Putnam’s metasemantic theory plays a part in explaining the epistemic grounds for first-order essentialist claims. Essentialists thus have good reason to invoke the theory as part of the overall defence of their view. However, we must also take care not to overstate the point. Salmon also argues that it is possible to reach essentialist conclusions within alternative metasemantic frameworks (186–9). For instance, one can derive the conclusion—that necessarily, water is H_20—by replacing (i) with a broadly Fregean metasemantic theory, whereby the reference of ‘water’ is fixed by a rigidified description. It follows, then, that causalist metasemantics isn’t the only route to explaining the necessary *a posteriori.*

I do not wish to dispute this.[113] Even if we grant that the phenomenon of necessary *a posteriori* truth is explicable in multiple metasemantic frameworks, the point still stands that a proper explanation, against the historical backdrop of widespread skepticism towards the necessary a posteriori, will presuppose an *inflationary* understanding of metasemantics. To explain the joint necessity and *a posteriori* status of ‘water is H_20’, one must explain how reference for ‘water’ is *rigid* and *dependent on external features of the subject’s environment* (e.g. the hidden structure of water). And to that end, one must say more about reference than what is permitted by the deflationist’s framework.

[112] It is also possible to explain the *necessary a posteriori* status of <water is H_20> by simply conjoining the first-order explanation of its necessity with the commonsense claim of its *a posteriority.* Such an explanation would be entirely adequate for most purposes. However, it would not constitute a satisfying response to the historical opponents of the necessary a posteriori. To defend the *necessary a posteriori* against the historical critics without begging the question, one needs an explanation of how we can secure reference to substances and kinds without knowing, *a priori,* what must be true of them. This explanation is offered by Putnam’s metasemantic theory. Moreover, this kind of explanation is unavailable to the deflationist.

[113] However, a case can be made that realism about essences sits much better with realism about the grounds of reference—i.e. that reference is grounded in real causal relations and objective microsimilarities between samples. It would take me too afield to make the case here, but it is for this reason that I’ve centred the discussion around Putnam’s theory.

5.5 Expressivism

If the above two points are on the right track, then so we see how inflationary concepts of truth and reference can corroborate the justification of the first-order beliefs that concern realistic domains. They accomplish this by explaining how we manage (in those cases) to represent a world that is not of our own making. When a belief concerns real entities and kinds, its semantic features are explicable (in part) by how the world is carved.

But just as inflationary concepts of truth and reference can aid us in investigating real entities and kinds, they may also serve when the chosen subject matter isn't robustly real. Say we are exploring a topic where the entities, kinds, and properties aren't 'carved by the world'; they are, in some sense, projections of our non-representational attitudes. Perhaps we are contemplating whether factory farming is wrong, and we have assumed that the property of *wrongness* is not explicable along traditional realist lines.[114] In that case, we would certainly deny that reference for 'wrong' is determined even partly by the hidden structure of a mind-independent property (*à la* Putnam for 'water'). Instead, we would presumably explain the semantic features of 'wrong' by adverting to our non-representational attitudes—e.g. desires, preferences, plans, and the like. However, as Taylor (2020) argues, even saying this much will likely court the ideology of an inflationary conception of truth.

The claim that metaethical irrealism invites *inflationism* will probably surprise those with a historical view of the metaethics literature. Traditionally, metanormative realism's most prominent rival has been metanormative expressivism—i.e. the claim that moral thought and talk function differently from ordinary descriptive/representational thought and talk. Rather than tracking a world of mind-independent facts, moral language functions, according to the expressivist, to express the speaker's evaluative attitudes, and moral judgments are (at least partly) constituted by those evaluative attitudes (Blackburn 1993, 1998; Gibbard 1990; Ridge 2014). However, despite these doctrines, there is overwhelming linguistic evidence that moral sentences and judgments are susceptible to truth and falsity.[115] So to square their theories with the linguistic data, expressivists have frequently allied their view with the deflationary view of truth (Blackburn 1998, 2008, 2010; Price 1994). The hope was that deflationism would allow for

[114] I caution the reader not to get too irked by this assumption. For our purposes, the main point of this section need not hang on the expressivist/projectivist framework's application to *morality* specifically. It suffices that there be *some* application for the framework—whether it be moral, aesthetic, or something else.

[115] The *locus classicus* for the kind of argument is Geach (1965). There is also the observation that moral sentences are embeddable in truth attributions and that the truth concept's generalization function is no less expedient when morality is concerned (Horwich 1998: 84).

moral truth without forfeiting the idea that moral statements function to express evaluative attitudes.

However, recent decades have shown that the alliance between expressivism and deflationism was beset with serious problems. As Dreier (2004) argues, the expressivists' motive for adopting truth deflationism also invites a series of other deflationist concessions. Accommodating the evidence of ordinary discourse requires them to commit to moral propositions, properties, and facts—all of which can be accounted for by various deflationary theories (Dreier 2004: 26). Given these deflationist commitments, the expressivist can agree with the realist on matters of truth-conditional semantics and first-order moral fact. They can agree, for instance, that *'factory farming is wrong' is true just in case factory farming is wrong,* and that '*wrong*' applies to all and only the *wrong* actions. They may even maintain (if they're so inclined) that factory farming *is,* in fact, *wrong*—that *it is true* that factory farming is wrong. It is worthwhile for the expressivist to agree to such things since it helps reconcile their view with orthodox semantics and ordinary moral practice. However, if *deflationism* is the backdrop that facilitates this reconciliation, then a worry emerges: at some point, it becomes difficult to see how expressivism is any different from realism.

This is the 'problem of creeping minimalism' (Dreier 2004: 28). The challenge is to explain how metanormative expressivism differs from realism, given that the two can make identical pronouncements on the truth-conditional semantic facts for moral thoughts and utterances. As I see it, the burden of this challenge lies primarily on the expressivists since they purport to offer a genuine alternative to realism.

There are numerous ways to try to meet this challenge (e.g. Blackburn 2007; Dreier 2004; Dunaway 2016). Since it is not my task to solve the problem here, I can spare much of the detail and skip straight to an observation made by Taylor (2020). In short, once the expressivist analyzes moral thought and talk in terms of truth conditions and reference, and reserves the right to mimic the realist in their first-order moral and metaphysical opinions, they leave themselves very little room to distinguish their position—*except for one remaining point:* the *metasemantics* of moral thought and talk. And indeed, this is what we generally find. Dreier (2004), for instance, locates the distinction between expressivism and realism in the two positions' respective explanations of belief-content attributions. Take the fact that *Graham believes that factory farming is wrong.* According to Dreier, the realist sees this fact as explicable by the external relations that Graham bears to the property of *wrongness* (2004: 41). Whereas for the expressivist, the explanation will reside in the facts about Graham's desires, plans, and preferences concerning factory farming (41).[116] Dunaway (2016) offers another variation of this

[116] On the face of it, Dreier's solution concerns belief attributions, not truth conditions. However, as Taylor (2020) argues, the most plausible reading of his solution is that the realism-expressivism divide is to be located in their respective explanations of the *content* of moral beliefs, where 'content' is read in a way that includes truth conditions (125–7).

theme. In his account, the realist explains the reference-fixation of 'wrong' by imputing Lewisian eliteness to the property of *wrongness.* (A property is *elite* insofar as its instances are objectively similar, and this fact explains its aptitude for semantic and epistemic access (2016: 249; see §4.5.2).) In contrast, the expressivist (according to Dunaway) denies that *wrongness* is elite; instead, they explain our semantic access to wrongness without presupposing objective similarities amongst its instances. Crucially, the realist and expressivist need not disagree over *what* we refer to with our use of moral language and concepts. The difference lies squarely in *how* the reference and truth-conditional facts are explained (2016: 261).

There is much more to be said to flesh out these proposals, but most of it will be irrelevant to our purposes here. For us, the crucial point is the one made by Taylor (2020): that each of these proposals draw the realist-expressivist divide precisely within the explanations of moral content, including truth conditions and reference.[117] But as Taylor (2020) argues, this strategy for solving the problem of creeping minimalism is inconsistent with the deflationary conception of truth. The reason for this should now be familiar. Deflationism entails that the facts of truth conditions and reference are trivial in the sense that they are not susceptible to any explanation beyond the DS and RDS schemas (2020: 105–19). However, explanations of this kind are precisely what these proposed solutions offer (121–7). They each locate the difference between expressivism and realism in the ways that they explain truth conditions and reference. So to make sense of expressivism along these lines, we must presuppose an inflationary understanding of truth.

No doubt, this conclusion will be contentious. (There are other ways of interpreting expressivism; c.f. Rosen 1998: 387–8; Schroeder 2008: 33). But let's suppose that it is on the right track. In that case, understanding our first-order moral judgments through an expressivist lens will require an inflationist backdrop. By itself, this would already be a remarkable result. It would reveal an inflationary function for the truth concept in our metaethical theorizing: the concept is key to making sense of a fundamental metaethical distinction.

However, if we have already supposed this much, then we can press the point further. Not only would the concept of truth serve an inflationary function in our metaethical theorizing, but it could also serve an inflationary role in our first-order ethical inquiry.

The argument for this is much like the case of scientific realism in §IV. If it can be argued that a meta-level theory of the nature of a domain of discourse (i.e. realism vis-à-vis scientific discourse; expressivism vis-à-vis ethical discourse) differentially sanctions specific methodologies, rules of inference, or epistemic practices for first-order inquiry in that domain, then anything that supports the higher-level theory can translate downwards as part of the

[117] Taylor writes that "what distinguishes the realist from the expressivist is not any 'surface-level' or 'first-order' feature of moral talk and thought—nothing to do with truth or assertion of the expression of properties… Rather what the difference between expressivism and realism consists in is something going on *below* the surface—*explanations* of that thought and talk, i.e. explanations of moral content." (2020: 127–8)

justification for the first-order beliefs that are obtained by those methodologies, rules of inference, or epistemic practices. If, moreover, the meta-level theory is supported by an inflationary conception of truth, then it follows that inflationism will have some part to play in the overall justification for the first-order beliefs thus obtained. So the question now is whether the meta-level theory at issue—expressivism—would have any impact on how we investigate the moral domain. In other words, would a belief in expressivism change how we think about first-order ethical issues (e.g. whether factory farming is wrong)?

Once again, I must be much briefer than the subject deserves. Still, I trust that there is widespread sympathy for the claim that an expressivist backdrop would affect ethical opinion and practice. Blackburn ends his (1998) with a poignant note on how the quasi-realist expressivist sees ethical practice:

> So what is the right method of ethics? … Remember that for quasi-realism, an ethic is the propositional reflection of the dispositions and attitudes, policies and stances, of people. The virtues of a system of ethics are simply (and exactly) the virtues of the people who live it. ... What we need to do is to make our responses mature, imaginative, cultured, sympathetic, and coherent... We stand on our own feet, and our feet are human feet. This is how it is, and how it must be. (1998: 310)

Unlike the realist, the expressivist's moral opinions are not beholden to the patterning of some non-natural property (Bedke 2020). Nor need they conform their views to the distribution of any one natural property. For this reason, they can shrug off certain skeptical challenges that cast doubt on our sensitivity to a mind-independent moral realm (e.g. Street 2006). Instead, expressivism licenses us to conduct our ethical inquiry from *within* our own given moral outlooks. If moral properties are projections of our evaluative attitudes, then the right methods for exploring this realm are ultimately the ones that best reflect our own moral sensibilities.[118]

5.6 Conclusion

Returning to our main theme, the purpose of these examples is to illustrate a recurring pattern of inquiry. What matters to us is less so the examples and more so the pattern itself.

To reiterate, we begin the pattern with an initial investigation into the states of the word, using first-order sentences and first-order thoughts. But then, as part of our endeavour to *justify* a

[118] There are many other, more-concrete ways in which metaethics can differentially impact first-order ethics. For instance, McGrath (2011) discusses how the various metaethical views bear on the appropriateness of pure testimonial deference—i.e. basing one's moral opinions solely on the testimony of others. For another example, Ayars (2021) argues that her version of expressivism rules out rational egoism.

given first-order claim, we semantically ascend and engage in a higher-order mode of reflection. In this reflection, we entertain and theorize the *grounds* through which we pick out the subject matter of our first-order thoughts. That is, we justify a metasemantic theory as to how the truth conditions of our first-order thoughts are determined. Next, our metasemantic theory plays a key role in justifying a methodology for investigating the given topic. In other words, the means through which we semantically access the facts vindicate a method for knowing those facts. Finally, we appeal to this methodology to justify our original first-order claim.

Throughout this chapter, we have found several reasons why theorists concerned with truth should be especially interested in this pattern. Chief among them is that it opens up a schism between the deflationary and inflationary theories of truth regarding their consequences for how we investigate the world. As I have argued, it is possible to embrace this pattern as an inflationist, but it is not possible as a deflationist. For the deflationist, it is misguided to uphold any non-trivial theory about the grounds for the truth-conditional semantic facts; *ipso facto,* such theories shouldn't play any role in justifying our views about the world. *Pace* much of the truth literature, deflationism is much more versatile than is ordinarily supposed. However, one thing that deflationism absolutely cannot do is allow for the concept of truth to be used to report on substantive language-world or mind-world relations. And in the line of reasoning just outlined, the concept of truth is essentially used to do just that.

Besides these claims about the extant debate, this situation also opens up the possibility of a novel way to challenge deflationism. If one can argue that the pattern is non-empty—that is, *if* a theory of metasemantic determination plays some crucial role in justifying a claim about the world—then one will have shown that deflationism is missing something important.

In this chapter, I have given a preliminary case for motivating this challenge. I did so by offering inflationary interpretations of two dialectical developments, one from the philosophy of science and one from metaethics. However, it is worth reiterating that the case against deflationism does not rest on any one application. It is enough that there are *some* instances of inquiry that follow the general pattern. And I think it is reasonable to expect to find others, especially on topics that primarily interest philosophers. Symptomatic of this pattern is whenever a first-order dispute becomes embroiled in a higher-order controversy over the metasemantics of the initial terms used in the debate. I trust that this symptom is recognizable across many philosophical sub-disciplines.

Finally, I must emphasize that the aims of this chapter are not wholly negative. If I am right about the applications, then a positive view emerges on the role of truth in first-order inquiry. According to my positive account, the concept of truth (and the predicate 'is true') are invaluable when ascribed to explicit thoughts or sentences because they facilitate semantic ascent and allow us to reflect on the conditions under which our thoughts and sentences are true. (This much is in agreement with Rattan.) This higher-order reflection further enables us to entertain and justify metasemantic theories for our first-order representations; it allows us to represent the grounds for semantic determination. Furthermore, I claim that this reflection can be valuable to first-order inquiry precisely when it plays a role in our broader efforts to justify our first-order beliefs and claims.

So, in short, what *good* is the concept of truth? One answer—the one I have been suggesting—is that the concept of truth is one of the tools that allows us to ask *why* our representations represent what they do. And it is a good thing that we have such a tool because it opens certain avenues for investigating the world.[119]

[119] The positive view requires a final clarificatory note. It is not part of the position advocated here that the concept of truth's *raison d'être* is to represent semantic determinants. I only claim that it is *a* function of the truth concept, not that it is *the* function for which the concept evolved. To echo my earlier comments, it is possible that the concept exapted this function after evolving for other purposes. This claim alone would be sufficient to challenge deflationism, since deflationism entails that the concept of truth has *no* inflationary functions.

Conclusion

The kind of view advanced in this dissertation—an object-based correspondence theory of truth—has had a tumultuous relationship with the extant literature. Some writers (Glanzberg 2015) claim that it is a dominant thread in current thinking about truth and representation. Others (Horwich 1998; Williams 2020: 37) regard it as hopeless.

But before we can properly estimate its plausibility, much more must be done to specify its commitments. Clarifying the view, and positioning it against its competitors, has been a large part of my project here. For instance, chapter one tells a story of how the object-based correspondence theory *naturalizes* the philosophy of truth. By this, I mean that it takes a subject that formerly belonged solely to metaphysics (the nature of truth) and brings it into the fold of other, more scientific and specialized disciplines (semantics, metasemantics, and cognitive science). With naturalization comes a hallmark of maturation: the theory of truth gets divided into several sub-topics that are each the purview of specialists. And according to the object-based theory, each of these sub-disciplines informs the theory of truth.

Chapter one also draws out the implications of the theory for several philosophical controversies. Because of its structure, the object-based correspondence theory cannot be neutral on the nature of truth-bearers and the order of metasemantic explanation. For this reason, a full defence of the theory will also require a defence of its commitments on truth-bearers and the productivist approach to metasemantics. However, both of these undertakings are too large for one project. Although I make some headway on the second task in chapter four, I hope to give a more detailed examination of these topics in future research, using the framework laid out in this dissertation.

In chapter two, I contrast the theory with one of its most prominent rivals: the deflationist theory of truth. The main aims of this chapter are to recast the debate between deflationists and traditional inflationists as a debate over what it takes to explain reference, and to argue that there is space for a plausible middle ground. I call the middle-ground 'moderate inflationism'. Moderate inflationism allows for the theory of reference to be both *substantial* (i.e. non-deflationary) and work piecemeal.

Generally speaking, the deflationary theories of truth all imply that reference relations cannot be substantively explained or adequately theorized. For this reason, deflationists advocate for 'trivial' accounts of reference, whereby reference is supposed to be entirely captured by various trivializing schemas. Traditional inflationists, by contrast, have sought after reductionist accounts of reference, whereby reference is supposedly reduced to a uniform set of non-semantic relations. Both of these pictures have in common a 'top-down' approach. By this, I mean that they each presume to impose a broad picture of what reference must look like upon the metasemantic accounts for each kind of term. My moderate inflationary view, by contrast, does not take any such top-down approach. Instead, it permits the explanations of reference to vary, depending on what's appropriate for the representation in question. And this, it seems to me, is a significant virtue. Moderate inflationism sits well with a methodology for metasemantics that embraces specialization. It permits each field of inquiry (cognitive science, philosophy of mind, metaethics, etc.) to pursue their metasemantic questions according to their own explanatory aims.[120]

Chapter three elaborates the framework for theorizing about reference from a moderate inflationist perspective. It also continues to build on this theme of specialization. In the foreground, the principal concern of this chapter is the metaphysics of reference. On this score, I elaborate a *pluralist* account that draws on the metaphysics of functionalism (following Lynch 2009). But in line with this pluralism, I also advocate for an approach to metasemantics that permits a variety of explanatory aims, depending on the subject matter of the referring item in question. In other words, the pluralist metaphysics of reference underwrites a pluralist methodology for metasemantic inquiry.

As I argue in chapter one, an object-based correspondence theory must prioritize reference over truth in the order of metaphysical explanation. For this reason, it must be paired with a productivist approach to metasemantics (Simchen 2017) as opposed to the interpretationist approach of Davidson (1973, 1974, 1977) and Lewis (1974, 1975). The aims of chapter four are thus two-fold. First, I clarify the differences between the functionalist account of the previous chapter and metasemantic interpretationism. Distinguishing the two is key to maintaining the overall integrity of the object-based correspondence theory combined with a functionalist metaphysics of reference. Secondly, I give a presumptive case for favouring the functionalist-productivist view over interpretationism. Since the debate between productivism and interpretationism is too large and multifaceted for a single chapter, I must settle on a presumptive case.

Ultimately, the main argument *for* this conception of truth must lie in its applications. The picture deserves credence only to the extent that it offers a framework for making sense of other significant issues. In chapter five, I apply the theory to answer another question of broad philosophical interest: *what role does our conception of truth play in inquiry about the world?* My principal aim is to contrast the answers afforded by moderate inflationism and the deflationist conception of truth. (To this end, I must also contrast my theory with another inflationary theory given by Rattan 2016). In my view, there is a pattern of thought that can only

[120] In her (2007), Penelope Maddy defends disquotationalism (a particularly hardline version of deflationism) as part of her general defence of methodological naturalism. For Maddy, there is no 'first philosophical' perspective for the philosopher to criticize a scientific discipline from on high; philosophical inquiry must begin from *within* (as part of) scientific inquiry. If my diagnosis in the last paragraph is correct, then disquotationalism may actually be in tension with Maddy's second philosophical outlook. That is because disquotationalism ventures to reinterpret and revise the theories of content delivered by cognitive science (e.g. Shea 2018) in light of general philosophical claims about truth.

be facilitated by an inflationist conception of truth. I call it ‘representing semantic determinants’. Basically, an inflationary conception of truth can explain the value of truth-conditional metasemantic reflection, and the opportunities for inquiry that it affords, whereas a deflationary conception cannot. Since—as I argue—there are occasions (primarily in philosophy) where metasemantic reflection is valuable for inquiry about the world, it follows that the deflationary conception of truth is missing something important. In place of deflationism, the moderately-inflationary, object-based correspondence theory is perfectly suited to underwrite these applications of the truth concept.

As far as I know, chapter five’s argument strategy represents a fairly under-explored way of challenging deflationism. To make this strategy work, the central component is to demonstrate an example where truth-conditional metasemantic reflection makes some valuable contribution to some inquiry about the world. In this dissertation, I mention two examples that plausibly fit this pattern: one from the philosophy of science and one from metaethics. However, my case for these examples had to be fairly preliminary. A full discussion of either of them would be an entire project on its own. Nonetheless, this whole strategy opens up a line of research that I can continue to pursue in the future. Rather than framing this future project as centrally concerned with truth, it can be more directly concerned with the question: *what epistemic role does metasemantic reflection play in philosophical debates?* (Example: *what epistemic role did the causal theory of reference play in the defence of scientific realism and essentialism?*) A proper discussion of any such example can further support the theory of truth I developed here.

References

Adams, F. and Aizawa, K. (2017) ‘Causal Theories of Mental Content’, in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), <https://plato.stanford.edu/archives/sum2017/entries/content-causal/> accessed September 26 2022.

Armour-Garb, B. & Beall, J. C. (2005) ‘Deflationism: the Basics’, in B. Armour-Garb & J.C. Beall (ed.) *Deflationary truth*, 1–29. Chicago and La Salle: Open Court Publishing.

Austin, J. L. (2018) *How To Do Things With Words.* Edited by J.O. Urmson. Connecticut: Martino Fine Books. (Original work published 1962.)

Ayars, A. (2021) ‘Deciding for Others: An Expressivist Theory of Normative Judgment’, *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12800

Ball, D. (2017) ‘Review of Semantics, Metasemantics, Aboutness’, *Notre Dame Philosophical Reviews*. <http://ndpr.nd.edu/news/semantics-metasemantics-aboutness/> accessed September 26 2022.

Bedke, M. S. (2020) ‘A Dilemma for Non-Naturalists: Irrationality or Immorality?’ *Philosophical Studies*, 177(4): 1027–42.

Blackburn, S. (1993) *Essays in Quasi-Realism*. Oxford: OUP.

Blackburn, S. (1998) *Ruling Passions.* Oxford: OUP.

Blackburn, S. (2006) ‘The Semantics of Non-factualism, Non-cognitivism, and Quasi-realism’, in M. Devitt & R. Hanley (eds.), *The Blackwell guide to the Philosophy of Language*, 244–52. Oxford: Blackwell Publishing Ltd.

Blackburn, S. (2007) ‘Anti-Realist Expressivism and Quasi-Realism’, in D. Copp (ed.), *The Oxford Handbook of Ethical Theory*, 146–62. Oxford: OUP.

Blackburn, S. (2010) ‘Truth, Beauty and Goodness’, in R. Shafer-Landau (ed.), *Oxford studies in Metaethics*, *5*, 295–314. Oxford: OUP.

Boghossian, P. (1989) ‘The Rule-Following Considerations’, *Mind,* 98/392: 507–549.

Boyd, R.N. (1983) ‘On the Current Status of the Issue of Scientific Realism’, in C.G. Hempel, H. Putnam, & W.K. Essler (eds.) *Methodology, Epistemology, and Philosophy of Science*, *45–90.* Springer, Dordrecht.

Brandom, R. (1996) ‘The Significance of Complex Numbers for Frege's Philosophy of Mathematics’, *Proceedings of the Aristotelian Society*, 96: 293–315.

Burgess, A. (2011) ‘Mainstream Semantics + Deflationary Truth’, *Linguistics and philosophy*, 34/5: 397–410.

Burgess, A. and Sherman B. (2014) ‘A Plea for the Metaphysics of Meaning’, in A. Burgess & B. Sherman (eds.) *Metasemantics: New essays on the foundations of meaning*, 1–16. Oxford: OUP.

Button, T. & Walsh, S. (2018) *Philosophy and Model Theory*. Oxford: OUP.

Chomsky, N. (1995) ‘Language and Nature’, *Mind*, 104/413: 1–61.

Cummins, R. (1997) ‘The Lot of the Causal Theory of Mental Content’, *The Journal of Philosophy*, 94/10: 535–42.

David, M. (2018) ‘The Correspondence Theory of Truth’ in M. Glanzberg (ed.) *The Oxford Handbook of Truth*, 238–58. Oxford: OUP.

Davidson, D. (1967) ‘Truth and Meaning’*, Synthese*, 17/3: 304–23.

Davidson, D. (1973) ‘Radical Interpretation’, *Dialectica*, 27: 313-28.

Davidson, D. (1974) ‘Belief and the Basis of Meaning’, *Synthese*, 27/3: 309–23.

Davidson, D. (1977) ‘Reality Without Reference’, *Dialectica*, 31/3–4: 247–58.

Davidson, D. (1979) ‘Quotation’, *Theory and Decision*, 11/1: 27–40.

Davidson, D. (1983) ‘A Coherence Theory of Truth and Knowledge’, reprinted in Davidson (2001) *Subjective, intersubjective, objective*, 137–57. Oxford: OUP.

Davidson, D. (1984) *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press.

Davidson, D. (1990) ‘The Structure and Content of Truth’, *The Journal of Philosophy*, 87/6: 279–328.

Davidson, D. (1991) ‘Three Varieties of Knowledge’, in A. Phillips Griffiths (ed.) *A.J. Ayer Memorial Essays: Royal Institute of Philosophy Supplements*, *30*, 153–66. Cambridge: Cambridge University Press.

Dickie, I. (2015) *Fixing reference*. Oxford: OUP.

Donnellan, K. S. (1966) ‘Reference and Definite Descriptions’, *The Philosophical Review*, 75/3: 281–304.

Dreier, J. (2004) ‘Meta-ethics and the Problem of Creeping Minimalism’, *Philosophical Perspectives*, 18/1: 23–44.

Dretske, F. (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT/Bradford Press.

Dunaway, B. (2016) ‘Expressivism and Normative Metaphysics’, R. Shafer-Landau (ed.) in *Oxford Studies in Metaethics*, *vol. 11,* 241–64. Oxford: OUP.

Edwards, D. (2018) *The Metaphysics of Truth*. Oxford: OUP.

Feyerabend, P. K. (1962) ‘Explanation, Reduction, and Empiricism’, in H. Feigl and G. Maxwell (ed.) *Scientific Explanation, Space, and Time. Minnesota Studies in the Philosophy of Science, Volume III*, 28–97. Minneapolis: University of Minneapolis Press.

Field, H. (1972) ‘Tarski's Theory of Truth’, *The Journal of Philosophy*, 69/13: 347–75.

Field, H. (1978) ‘Mental Representation’, *Erkenntnis*, 13/1: 9–61.

Field, H. (1994a) ‘Deflationist Views of Meaning and Content’, *Mind*, 103/411: 249–85.

Field, H. (1994b) ‘Disquotational Truth and Factually Defective Discourse’, *The Philosophical Review*, 103/3: 405–52.

Field, H. (2005) 'Postscript to "Deflationist Views of Meaning and Content"', in B. Armour-Garb & J. C. Beall (eds.) *Deflationary truth,* 92–110. Chicago and La Salle: Open Court Publishing.

Field, H. (2008) *Saving Truth From Paradox*. Oxford: OUP.

Fodor, J.A. (1968) *Psychological Explanation*. New York: Random House.

Fodor, J.A. (1974) 'Special Sciences (Or: The Disunity of Science as a Working Hypothesis)', *Synthese*, 28/2: 97–115.

Fodor, J. A. (1975) *The Language of Thought*. New York: Crowell.

Fodor, J. A. (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1990) *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. A. (1998) *Concepts: Where Cognitive Science Went Wrong*. Oxford: OUP.

Fodor, J. A. (2008) *LOT 2: The Language of Thought Revisited*. Oxford: Clarendon Press.

Frege, G. (1879) *Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought*. Partially reprinted in M. Beaney (ed.) (1997) *The Frege Reader*, 47–78. Oxford: Blackwell Publishing Ltd.

Frege, G. (1948) 'Sense and Reference', *The Philosophical Review*, 57/3: 209–30.

Frege, G. (1956) 'The Thought: A Logical Inquiry', *Mind*, 65/59: 289–311.

Geach, P. T. (1965) 'Assertion', *The Philosophical Review*, 74/4: 449–65.

Gibbard, A. (1990) *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Glanzberg, M. (2015) 'Representation and the Modern Correspondence Theory of Truth', in S.

Gross, N. Tebben, & M. Williams (eds.) *Meaning without Representation: Essays on Truth, Expression, Normativity, and Naturalism*, 81–102. Oxford: OUP.

Glüer, K. (2018) ‘Interpretation and the Interpreter’, in D. Ball and B. Rabern (eds.) *The Science of Meaning: Essays on the Metatheory of Natural Language Semantics*, 226–52. Oxford: OUP.

Gupta, A. (1993) ‘A Critique of Deflationism’, *Philosophical Topics*, 21/2: 57–81.

Hanks, P. (2009) ‘Recent Work on Propositions’, *Philosophy Compass*, 4/3: 469–86.

Haslanger, S. (2020) ‘Going On, Not in the Same Way’, in A. Burgess, H. Cappelen, & D. Plunkett (eds.) *Conceptual Engineering and Conceptual Ethics*, 230–60. Oxford: OUP.

Heim, I. & Kratzer, A. (1998) *Semantics in Generative Grammar.* Malden, MA: Blackwell Publishing Ltd.

Henkin, L. (1949) ‘The Completeness of the First-Order Functional Calculus’, *Journal of Symbolic Logic*, 14/3: 159–66.

Henkin, L. (1950) ‘Completeness in the Theory of Types’, *Journal of Symbolic Logic*, 15/2: 81–91.

Horisk, C. (2008) ‘Truth, Meaning, and Circularity’, *Philosophical Studies*, 137/2: 269–300.

Horwich, P. (1998a) *Truth*. Oxford: OUP.

Horwich, P. (1998b) *Meaning*. Oxford: OUP.

Horwich, P. (2005) ‘Reference’, in B. Armour-Garb & J. C. Beall (eds.) *Deflationary truth*, 184–98. Chicago and La Salle: Open Court Publishing.

Kaplan, D. (1978) ‘Dthat’, *Syntax and Semantics* 9: 221–43. Reprinted in P. French, T. Uehling, & H. Wettstein (eds.) *Contemporary Perspectives in Philosophy of Language*, 383–400 Minneapolis, MN: University of Minnesota Press.

Kaplan, D. (1989a) ‘Afterthoughts’, in J. Almog, J. Perry, and H. Wettstein (eds.) *Themes From Kaplan*, 565–614. Oxford: OUP.

Kaplan, D. (1989b) ‘Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and other Indexicals’, in J. Almog, J. Perry & H. Wettstein (eds.) *Themes From Kaplan*, 581–63. Oxford: OUP.

King, J. C. (2007) *The Nature and Structure of Content*. Oxford: OUP.

King, J. C. (2014a) ‘What Role do Propositions Play in our Theories’, in J. C. King, S. Soames & J. Speaks (eds.) *New Thinking About Propositions*. 5–8. Oxford: OUP.

King, J. C. (2014b) ‘The Metasemantics of Contextual Sensitivity’, in A. Burgess & B. Sherman (eds.) *Metasemantics: New Essays on the Foundations of Meaning*, 97–118. Oxford: OUP.

King, J. C., Soames, S., & Speaks, J. (2014) *New Thinking About Propositions*. Oxford: OUP.

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

Kripke, S. (1982) *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Oxford: Blackwell Publishing Ltd.

Kuhn, T. (1962) *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Leeds, S. (1978) ‘Theories of References and Truth’, *Erkenntnis*, 13/1: 111–29.

Leeds, S. (1995) ‘Truth, correspondence, and success’, *Philosophical Studies*, 79/1: 1–36.

Leeds, S. (2000) ‘A Disquotationalist Looks at Vagueness’, *Philosophical Topics*, 28/1: 107–28.

Lewis, D. (1970) ‘How to Define Theoretical Terms’, *The Journal of Philosophy*, 67/13: 427–46.

Lewis, D. (1972) ‘Psychophysical and Theoretical Identifications’, *Australasian Journal of Philosophy*, 50/3: 249-258.

Lewis, D. (1974) ‘Radical Interpretation’, *Synthese*, 27/3–4: 331–44.

Lewis, D. (1975) ‘Languages and Language’, reprinted in A.P. Martinich & D. Sosa (eds.) (2013) *The Philosophy of Language: Sixth Edition*, 682–700. Oxford: OUP.

Lewis, D. (1984) ‘Putnam's Paradox’, *Australasian Journal of Philosophy*, 62/3: 221–36.

Lewis, D (1986) *On the Plurality of Worlds.* Oxford: Blackwell Publishing Ltd.

Lewis, D. (1969) *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

Lipton, P. (2004) *Inference to the Best Explanation (Second Edition)*. London and New York: Routledge.

Loewer, B. (1987) ‘From Information to Intentionality’, *Synthese*, 70/2: 287–317.

Lynch, M. P. (2009) *Truth as One and Many*. Oxford: OUP.

Maddy, P. (2007) *Second Philosophy: A Naturalistic Method*. Oxford: OUP.

McGrath, M. (1997) ‘Weak Deflationism’, *Mind*, 106/421: 69–98.

McGrath, S. (2011) ‘Skepticism about moral expertise as a puzzle for moral realism, *The Journal of Philosophy*, 108/3: 111–37.

McLaughlin, B. (2006) ‘Is Role-Functionalism Committed to Epiphenomenalism?’, *Journal of Consciousness Studies*, 13/1–2: 39–66.

Millikan, R. G. (1984) *Language, Thought, and Other Biological Categories: New Foundations For Realism*. Cambridge, MA: MIT press.

Moore, G. E. (1899) ‘The Nature of Judgment’, *Mind*, 8/30: 176–93.

Moore, G. E. (1953) *Some Main Problems of Philosophy.* London: George Allen & Unwin.

Moore, G. S. (2020) ‘Theorizing About Truth Outside of One’s Own Language’, *Philosophical Studies*, 177/4: 883–903.

Moore, G. S. (2022) ‘Between Deflationism and Inflationism: A Moderate View on Truth and Reference’, *The Philosophical Quarterly*, 72/3: 673–94.

Neander, K. (2017) *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

Neander, K. & Schulte, P. (2022) 'Teleological Theories of Mental Content', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), <https://plato.stanford.edu/archives/sum2022/entries/content-teleological/> accessed September 29, 2022.

Palmira, M. (2018) 'Towards a Pluralist Theory of Singular Thought', *Synthese*, 195/9: 3947–74.

Pagin, P. & Westerståhl, D. (2010a) 'Compositionality I: Definitions and Variants', *Philosophy Compass*, 5/3: 250–64.

Pagin, P. & Westerståhl, D. (2010b) 'Compositionality II: Arguments and Problems', *Philosophy Compass*, 5/3: 265–82.

Papineau, D. (1984) 'Representation and Explanation', *Philosophy of Science*, 51/4: 550–72.

Papineau, D. (1993) *Philosophical Naturalism*. Oxford: Blackwell.

Price, H. (1994) 'Semantic Minimalism and the Frege point', in S.L. Tsohatzidis (ed.) *Foundations of Speech Act Theory*, 140–63. London and New York: Routledge.

Putnam, H. (1973) 'Meaning and Reference', *The Journal of Philosophy*, 70/19: 699–711.

Putnam, H. (1975) 'The Meaning of "Meaning"', in K. Gunderson (ed.) *Minnesota Studies in the Philosophy of Science VII: Language, Mind, and Knowledge*, 139–93. Minneapolis: University of Minnesota Press.

Putnam, H. (1981) *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. (1948) 'On What There Is', *The Review of Metaphysics*, 2/1: 21–38.

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

Quine, W. V. O. (1986) *Philosophy of Logic*. Cambridge, MA: Harvard University Press.

Quine, W. V. O. (1987) *Quiddities*. Cambridge, MA: Harvard University Press.

Rasmussen, J. (2014) *Defending the Correspondence Theory of Truth*. Cambridge: Cambridge University Press.

Rattan, G. (2010) ‘Metarepresentation and the Cognitive Value of the Concept of Truth’, in C. D. Wright & N. J. L. L. Pedersen (eds.) *New Waves in Truth*, 139–56. London: Palgrave Macmillan.

Rattan, G. (2016) ‘Truth Incorporated’, *Noûs*, 50/2: 227–58.

Reimer, M. (1991) ‘Demonstratives, Demonstrations and Demonstrata’, *Philosophical Studies* 63/2: 187–202.

Ridge, M. (2014) *Impassioned Belief*. Oxford: OUP.

Rosen, G. (1998) ‘Blackburn’s Essays in Quasi-Realism’, *Noûs*, 32/3: 386–405.

Russell, B. (1919) ‘Descriptions’, reprinted in A.P. Martinich & D. Sosa (eds.) (2013) *The Philosophy of Language: Sixth Edition*, 114–120. Oxford: OUP.

Russell, B. (1997) *The Problems of Philosophy.* Oxford: Oxford University Press.

Russell, B. (2007) ‘The Philosophy of Logical Atomism’, in R. C. Marsh (ed.) *Logic and Knowledge*, 175–282. Nottingham: *Spokesman*.

Russell, B. (2020) *Principles of Mathematics*. London and New York: Routledge.

Salmon, N. (2005) *Reference and Essence (second edition)*. New York: Prometheus Books.

Schaffer, J. (2012) ‘Grounding, Transitivity, and Contrastivity’, in F. Correia & B. Schnieder (eds.) *Metaphysical Grounding: Understanding the Structure of Reality*, 122–38. Cambridge: Cambridge University Press.

Schroeder, M. (2008) *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: OUP.

Shea, N. (2018) *Representation in Cognitive Science.* Oxford: OUP.

Sher, G. (2015) ‘Truth as Composite Correspondence’ in T. Achourioti *et al.* (eds.) *Unifying the Philosophy of Truth*, 91–210. Springer, Dordrecht.

Sher, G. (2016) *Epistemic Friction: An Essay on Knowledge, Truth, and Logic*. Oxford: OUP.

Sider, T. (2010) *Logic for Philosophy*. Oxford: OUP.

Sider, T. (2013) *Writing the Book of the World.* OUP: Oxford.

Simchen, O. (2017) *Semantics, Metasemantics, Aboutness*. Oxford: OUP.

Soames, S. (1999) *Understanding Truth*. Oxford: OUP.

Soames, S. (2010) *What is Meaning?*. Princeton: Princeton University Press.

Stampe, D. (1977) 'Toward a Causal Theory of Linguistic Representation', *Midwest Studies in Philosophy*, 2/1: 42–63.

Strawson, P. F. (1950) 'On Referring', *Mind*, 59/235: 320–44.

Street, S. (2006) 'A Darwinian Dilemma for Realist Theories of Value', *Philosophical Studies*, 127/1: 109–66.

Tarski, A. (1944) 'The Semantic Conception of Truth and the Foundations of Semantics', reprinted in A.P. Martinich & D. Sosa (eds.) (2013) *The Philosophy of Language: Sixth Edition*, 375–97. Oxford: OUP.

Taylor, D. E. (2017) 'Deflationism and Referential Indeterminacy', *Philosophical Review*, 126/1: 43–79.

Taylor, D. E. (2020) 'Deflationism, Creeping Minimalism, and Explanations of Content', *Philosophy and Phenomenological Research*, 101/1: 101–29.

Taylor, K. A. (2019) *Meaning Diminished: Toward Metaphysically Modest Semantics.* Oxford: OUP.

Thomasson, A. L. (2014a) 'Deflationism in Semantics and Metaphysics', in A. Burgess & B. Sherman (eds.) *Metasemantics: New Essays on the Foundations of Meaning*, 185–213. Oxford: OUP.

Thomasson, A. L. (2014b). *Ontology Made Easy*. Oxford: OUP.

Van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford: Clarendon Press.

Williams, J. R. G. (2007) ‘Eligibility and Inscrutability’, *The Philosophical Review*, 116/3: 361–99.

Williams, J. R. G. (2020) *The Metaphysics of Representation.* Oxford: OUP.

Williams, M. (1999) ‘Meaning and Deflationary Truth’, *The Journal of Philosophy*, 96/11: 545–64.

Williamson, T. (2007) *The Philosophy of Philosophy.* Oxford: Blackwell Publishing.

Wittgenstein, L. (1961) ‘Notes on Logic’ in G. H. von Wright and G.E. Anscombe (eds.) *Notebooks 1914–1916*, 93–106. Oxford: Blackwell Publishing.

Wittgenstein, L. (2001) *Tractatus Logico-Philosophicus*. Translated by D.F. Pears & B.F. McGuinness. London and New York: Routledge.

Wright, C.J.G. (1992) *Truth and Objectivity*. Cambridge, MA: Harvard University Press.