HARLEQUIN SEMANTICS

by

MICHAEL JOHNSON

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of Ernest Lepore and Barry Loewer

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2011

ABSTRACT OF THE DISSERTATION

Harlequin Semantics

By MICHAEL DAVID JOHNSON

Dissertation Directors:

Ernest Lepore and Barry Loewer

> My great principle, as regards natural things, is that of Harlequin, Emperor of the Moon,... that it is always and everywhere in all things just like here. ~Leibniz

This dissertation is about Semantic Uniformity. Semantic Uniformity is the claim that what is true for some expressions is true for them all—at least, when it comes to semantics. In particular, I defend three claims in three chapters, in this order:

First, all simple linguistic expressions, and not just some, are non-descriptive. That is, their referents are not determined by fit with our beliefs.

Second, all simple linguistic expressions are rigid. Relative to each possible world, construed as a world of evaluation, every expression has one and the same referent.

These two claims point toward a general picture of why simple expressions refer to what they do. Simple expressions have their referents determined by the causal, informational, or lawful connections they bear to objects and properties in the world. Furthermore, this referential connection is ‘direct’ in that it doesn’t depend on the distribution of objects and properties at a world; simple expressions simply refer, and that explains their rigidity.

In the third chapter I defend the view that the same holds for complex expressions. Consensus has it that complex expressions derive their meanings from the meanings of their parts and how those parts are combined. But, I argue, a better account assumes that complex expressions have their referents determined by the causal, informational, or lawful connections they bear to objects and properties in the world, just like simple expressions.

In the background of all this is the Referentialist thesis, that there is no further aspect to meaning than reference. If Referentialism is accepted, then the argument of the dissertation is that we can have a "complete" theory of meaning simply by extending a direct causal theory of reference to all expressions.

ACKNOWLEDGEMENTS

Writing a dissertation requires a lot of planning. What did Kant say about that again?

> Making plans is often a luxuriant, boastful occupation of the mind by which a man gives himself the airs of creative genius, demanding what he cannot perform himself, censuring what he cannot do better, and proposing what he himself does not know where to look for.

Ah, yeah, that. Well, completing plans is often an arduous, humiliating occupation of the mind by which a man discovers how little he knows, fails to perform what he has demanded of himself, discovers that he has been bettered, and settles for what he can find instead of what he was looking for. And that 'a man' there is a particular indefinite. Particularly, it's me. Just as often though, completing plans is a fulfilling, collaborative engagement, where he learns what he was formerly ignorant of, becomes capable of doing more than formerly he could, builds on the work of his betters, and finds new avenues of exploration. Perhaps that happened in the dissertation; I should know, but I was busy at the time.

Among the people I'd like to thank for contributing to this project, sometimes in ways that were barely clinically significant and more frequently in ways that were life-saving are: Luvell Anderson, Marcello Antosh, Josh Armstrong, Jay Atlas, Pavel Davydov, Heather Demarest, Ophelia Deroy, Tom Donaldson, Richard Dub, Andy Egan, Jerry Fodor, Randy Gallistel, Thony Gillies, Gabriel Greenberg, Allison Hepola, Mike Hicks, Erik Hoversten, Henry Jackman, Alex Jackson, Ben Levinstein, Karen Lewis, Martin Lin, Bob Matthews, Zachary Miller, Ricardo Mena, Lisa Miracchi, Jennifer Nado,

Carlotta Pavese, Bryan Pickel, Ken Safir, Roger Schwarzchild, Andrew Sepielli Chung-chieh Shan, Barry Smith, Ted Sider, Will Starr, Matthew Stone, and Jenn Wang

The other people are the ones I forgot to include. And also my committee: Matthew Stone, who was never afraid to give harsh criticism, sound advice, and freely of his wine; Jason Stanley, who saw me through *two* dissertations, all while feeling good about being a gangster, and probably also a father; Jeff King, who gave me extensive comments multiple times on every draft, mere days after I sent it to him, even though I never once put page numbers on them as he requested; my co-Chair Barry Loewer, who knows a surprising lot about things he hasn't thought about for twenty years; and my co-Chair (and erstwhile co-author), Ernie Lepore, without whose no-nonsense approach I would probably never have finished.

I'd also like to thank (again, in case you didn't notice) my wife, Jenny Nado, who let me copy-paste my dissertation into hers so I didn't have to figure out the formatting. And also whom I love deeply and want to be with always. Jenny's stuff is spot-on so, if you like this dissertation, you'll love hers. And I'm not just saying that.

Finally, thanks are due to my parents, who raised me, paid my way through college, supported my dream of not having a real job, and still support me even after the dream has come true. Also, I suppose, my brother, who doesn't push me around, even though he's bigger than me and when I was bigger than him, I pushed him around. If that isn't love…

TABLE OF CONTENTS

Chapter 1

Descriptivism and General Terms

In this chapter, I want to argue that all general terms are non-descriptive. Here's the plan. First, I shall lay out the reasons why it would be nice if all general terms, and not just natural kind terms, were non-descriptive. These are the benefits that accrue to my view; this is what I gain should I establish it. Then, I'll lay out the evidence that Kripke and Putnam had for thinking that natural kind terms are non-descriptive and, subsequently, will show that these selfsame arguments extend to general terms generally. Finally, I'll consider objections, claim victory, and appropriate my prizes. So, on to the prizes.

1. The attractions of anti-descriptivism

The first is simply the **uniformity principle**. Kripke, Putnam, and Donnellan have succeeded in convincing most philosophers that names and natural kind terms have a non-descriptive metasemantics—that is, their semantic value is not determined by a set of properties speakers *associate* with them. The natural result has been for philosophers to accept some sort of causal account of representation, in lieu of the descriptive one. But this has a pretty clear consequence, as Stalnaker points out:

> If representation is essentially a causal relation, then *no* predicate, and *no* mental state, can represent in virtue of the intrinsic psychological properties of the person who is using the predicate, or who is in the mental state. Purely general properties may still be properties of things in the world, and representing such properties requires interaction with such things. [(1984) p. 21; emphasis added]

If we genuinely want to hold that names and natural kind terms have a causal metasemantics, we must either accept Stalnaker's conclusion, that every predicate has a causal metasemantics and thus no predicate has a descriptive metasemantics, or we must deny uniformity and say that representation is not essentially causal: it is causal here, but not over there.

Denying uniformity has its costs: for it is quite mysterious why the truth-value of a sentence like 'Bessy is a large cow' should depend, on the one hand, on the things, Bessy and cowhood, that 'Bessy' and 'cow' stand in a certain causal relation to, and also depend, on the other hand, on the thing, largeness, that 'large' bears an entirely different relation to, a relation of fit with the speaker's current mental state. If things are the same here as there, then we're done: we've discovered the nature of representation. If they're one way here and another there, then we'll have to say quite a lot about how all the different ways of representing are united, and how they can fit into the exact same explanations, while nevertheless differing so greatly.

The second benefit to wholesale anti-descriptivism is a solution to the **incommensurability problem**. One of Putnam's motivations in "The Meaning of 'Meaning'" and other work was to show how two individuals with different theories of the nature of, say, gold could both be talking about the same thing. If on both theories, gold is identified with the description "the stuff that plays the following role in the theory…" then both theories will be talking past one another—they might both be right, because they both identify 'gold' with different descriptions. But this same motivation extends to other (non-natural) kind terms: we may have different theories of the function of some ancient artifact, different theories of the etiology of some rock formation, and

different theories of the structure of some social practice (in, say, a foreign culture). If the theories are to be talking about the same things, then the terms used by the theories must not be identified with descriptions of the role the terms play in the theory. Thus, if only natural kind terms and not general terms generally are non-descriptive, then the specter of incommensurability will loom over theories in geology, anthropology, sociology, etc.

Finally, naturalness is somewhat of a bogus concept, at least as it is generally employed in the literature. One set of potential paradigms for natural kinds are substance terms like ‘gold’ and ‘water.’ To be gold is to be constituted of a certain stuff; to be water to be constituted of a different sort of stuff. A distinct sort of paradigm for natural kinds are the species. Never mind for the moment that ‘species’-talk only makes sense given the historical accidents of particular extinction events (if all the ancestors of all living things were alive today, there’d be no clear boundaries along which to draw species lines—this makes hay if you want to think of naturalness as a kind of joint carving). The point I want to press is that what it takes to be a tiger is not at all like what it takes to be gold. At the very least, a tiger’s etiology is important for its identity qua tiger. Martian “tigers”—animals with identical DNA to tigers but a causally unconnected evolutionary history—are not, we intuit, tigers. Duplicating a tiger is insufficient for tigerhood; one must also duplicate tiger etiology. However, not every property that is individuated along etiological lines makes it into the privileged set of “natural kinds”: duplicates of footprints, with distinct etiology, are no more footprints than Martian “tigers” are tigers. But somehow ‘tiger’ and ‘footprint’ are to be treated differently.

Yet a third paradigm for natural kinds are organs (some evidence for this claim: it is a posteriori yet necessary that hearts are for pumping blood). What it takes to be a heart

is different from what it takes to be gold, or to be a tiger. To be a heart is to have a certain function, a certain telos, to be *for* a particular purpose (i.e. pumping blood). It matters not what hearts are made of, or from whence they came. But again, not all functional kinds are traditionally counted "natural": artifacts are all individuated in terms of their functions. I don't even want to get in to what sort of kind *pain* might be, but I suspect it's not like gold, tigers, or hearts. And *red* may well be some sort of response-dependent property. There is a **disunity of naturalness**: sometimes so-called natural kinds are substance kinds, sometimes etiological kinds, sometimes functional kinds, sometimes yet other kinds; yet to maintain a natural/ non-natural divide it is necessary to assume that not all substance, etiological, or functional kinds are natural. Natural kinds don't form a natural kind.

When Quine introduced "natural kind" into the philosophical vocabulary, naturalness was supposed to be that quality of a property, whatsoever it was, that made it susceptible to induction (i.e. Quine was out to "solve" Goodman's new riddle of induction). I'm skeptical, to say the least, that Quine's notion of naturalness is Kripke's. If I eat 350 Big Macs (on separate occasions, controlling for confounding variables, etc.) and they each make me sick, then I'm justified in inferring that the next Big Mac I eat will make me sick. But nobody is inclined to think that 'Big Mac' is a natural kind term, in Kripke's sense.

I think that's enough to justify the project. If it turns out that all general terms, and not just natural kind terms, are non-descriptive, then we save semantic uniformity, solve the incommensurability problem for geology, sociology, anthropology, etc., and purge ourselves of an untenable distinction between what's "natural" and what isn't.

2. Kripke and Putnam against natural kind term descriptivism

Descriptivism with respect to natural kind terms says that such a term is descriptive, as used by a speaker or set of speakers A, iff its extension is determined by a set of properties (of individuals) that A associates with it:

> NK0. To every natural kind term "K", there corresponds a cluster of properties (of individuals), namely the family F of properties φ such that A believes 'all K's are φ'.
>
> NK1. An object x satisfies most or a weighted most of the properties in F if x satisfies "K" (as used by A).
>
> NK2. An object x satisfies "K" (as used by A) if x satisfies most or a weighted most of the properties in F.

The main argument that Kripke presents against the claim that natural kind terms satisfy condition (NK1) is what I'll call the Argument from Error. Chemical kinds are individuated by the sort of stuff that constitutes them and biological kinds by something like the structure of their DNA and their place in evolutionary history. But we humans are a limited sort, and often we must make do with identifying natural kinds by their macroscopically observable features, not their fundamental constitution, DNA, or history. That is, the observable features of a natural kind K might be the only features we associate with "K", and thus the only candidates for F. But it is always possible that we are in error with respect to all observable features of K, because of perceptual illusions, demons, magicians, false testimony, or other errors. In such cases, nothing satisfies most, or a weighted most, of the properties in F, and therefore (NK1) predicts that "K" will

have an empty extension. But this is the wrong result, as Kripke argues in the case of gold:

> Suppose there were an optical illusion which made [gold] appear to be yellow; but, in fact, once the particular properties of the atmosphere were removed, we would see that it is actually blue… Would there on this basis be an announcement in the newspapers 'It has turned out that there is no gold. Gold does not exist. What we took to be gold was not in fact gold.'?... It seems that there would be no such announcement. [(1980) p. 118]

And then again in the case of tigers:

> [W]e might… find out that tigers had none of the properties by which we originally identified them. Perhaps none are quadrupedal, none tawny yellow, none carnivorous, and so on; all these properties turn out to be based on optical illusions or other errors. [*Ibid*, p. 121]

The main argument against the claim that natural kind terms satisfy condition (NK2) may be called the Argument from Ignorance[1]. Just as we may be subject to illusion as epistemically fallible agents, so too we may suffer from a simple lack of knowledge. In particular, we may not possess enough knowledge about the properties possessed by members of a natural kind K to distinguish them from all non-members. If such a case is possible, then an object's satisfying the associated family F will not be a guarantee that it is a K, that is, (NK2) would be false. And such cases *are* possible:

> Suppose you are like me and cannot tell an elm from a beech tree. We still say that the extension of 'elm' in my idiolect is the same as the extension of 'elm' in anyone else's, viz., the set of all elm trees, and that the set of all beech trees is the extension of 'beech' in *both* of our idiolects. [Putnam, (1973) p. 704]

> Is it true that anything satisfying this description in the dictionary is necessarily a tiger? It seems to me that it is not. Suppose we discover an animal which, though having all external appearances of a tiger described here, has an internal structure completely different from that of a tiger… Let's say they were in fact very peculiar looking reptiles. Do we then conclude on the basis of this description that some tigers are reptiles? We don't." [Kripke, (1980) p. 120]

[1] No, not *that* Argument from Ignorance.

I should also remark that Putnam's famous Twin Earth case is also a version of the Argument from Ignorance: today, we know that water is H_2O, and this information excludes XYZ from the extension of 'water.' In 1750, we lacked the relevant chemistry, so the cluster of properties associated with 'water' included XYZ in their extension. But the meaning of 'water' never changed, so being in the extension of the cluster doesn't suffice for being in the extension of 'water,' that is, (NK2) is false for 'water.'

One thing to note about these arguments is that they rely crucially on the uniformity principle. Kripke and Putnam point out that there are *some* natural kinds for which we lack knowledge of either necessary conditions for their instantiation, or sufficient conditions for their instantiation. They nowhere argue that this is *always* the case, but they nevertheless conclude that *all* natural kind terms are non-descriptive. This is so, even though the Arguments from Ignorance and Error can't always be run. For instance, when we introduced 'ununseptium' into the language, we were in possession of necessarily necessary and sufficient conditions for being ununseptium, namely, being the element with 117 protons. This does not mean that 'ununseptium' is descriptive, for our epistemic situation with respect to some kind is irrelevant to how it has its content determined. The evidence that it is non-descriptive is first, that it is a natural kind term and second, that many other natural kind terms are non-descriptive.

Here then is our roadmap for determining whether other general terms besides natural kind terms are non-descriptive. First, we must identify the different classes of general terms—in addition to natural kind terms, for instance, there are functional, etiological, and structural kind terms. Then, we must show that some such terms in each class are non-descriptive (using the Arguments from Ignorance and Error), and that this is

not because of some special feature they have, that other members of their class do not. Thus, by uniformity, all members of these classes of terms will be non-descriptive. If the taxonomy is broad enough, and includes enough of the general terms in our language, a second application of the uniformity principle will get us that all general terms are non-descriptive (even those not properly subsumed by our taxonomic categories). There is of course no small danger in inferring from some to all, but no inquiry is perfect and, at the very least, these inferences will be buttressed by the more general considerations for non-descriptivity adduced in the introduction, above.

3. Non-natural kind terms: arguments from error

Let us turn here to the case of non-natural kinds. In this section, I'll introduce various kinds of kind terms, and go on to show how to construct Arguments for Error centered around them. The thesis we'll be concerned with here is GT1, that satisfying most or a weighted most of the properties a group of speakers associates with "G" is a *necessary condition* for being G:

> GT0. To every non-natural kind general term "G", there corresponds a cluster of properties (of individuals), namely the family F of properties φ such that A believes 'all G's are φ'.
>
> GT1. An object x satisfies most or a weighted most of the properties in F if x satisfies "G" (as used by A).

One class of non-natural kind general terms is the *functional kind terms*, which express properties of having a certain teleology: a function, use, or intended purpose. Objects, artifacts, and phenotypes can acquire a telos in two ways: first, an agent, such as a human, orangutan, or Martian, may intend a thing to fulfill a certain function; and

second, evolution may select for a phenotype that fulfills a certain function. Functional kinds of the first type (*artifact kinds*) include nests, hammers, maps, and shoes; and under the second heading we place hearts, lungs, livers, and flanges. Hearts differ greatly in their observable properties: color, size, shape, tissue profile, number of chambers; but they all share a function, to pump blood. Of note, they are all invariably possessed by organisms with a need for blood circulation, and this is what preserves genes for hearts in the gene pool. It is worth noting that some functional kinds are individuated both by their telos and how they achieve it: erasable pens are not pencils, though they play most roles pencils do; and were I to create an artifact perceptually indistinguishable from a microwave, that used gamma waves to heat food, it would not thereby *be* a microwave.

At one point in human history, we did not know the functions of most of our internal organs. It was a significant scientific discovery that hearts were *for* pumping blood. And even when it was discovered, many erred with Galen who might have been right with Harvey. Consider the following case:

> HEART: Fred lives in the 2nd Century C.E. If asked to state everything he believed about hearts (which he calls '*καρδιά*'), he would say that they are the seat of the soul, that their purpose is to create some but not all of the blood in the body ("bright blood" as opposed to "dark blood"), that they are red in color, and shaped roughly like contemporary Valentine's day cards. Quite unknown to Fred and his contemporaries, none of his beliefs regarding hearts is true. The seat of the soul is the pineal gland; hearts pump blood, but do not create it; they are green (but appear red because of an atmospheric illusion), and they are not what we now call "heart-shaped."

We may now ask, echoing Kripke: suppose that Fred's errors were discovered to him and his community. Would there on this basis be an announcement in the forum 'It has turned out that there are no hearts. Hearts do not exist. What we took to be hearts were not in fact hearts.'? It seems that there would be no such announcement. Neither

Fred nor his contemporaries take their beliefs about hearts to constitute a necessary condition for being a heart.

I understand that many will want to say, "but 'heart' is a natural kind term. This case tells us nothing we didn't already know." And though I am inclined to doubt that 'heart' is a natural kind term (or perhaps to doubt that concept of naturalness is clear enough to have an application), I admit that this is a reasonable position to hold. But the point here is that the same sort of reasoning can be applied to functional kind terms that are clearly not also natural kind terms, like artifact kind terms.

It might be thought that we *must* know the function of artifacts, and thus know necessary conditions for a thing's being this or that artifact, as it's we who give the function to them. But of course that's not true, and furthermore it's Earth-ist and alive-ist. Alien and long-dead civilizations produce artifacts too. So consider this case:

> BÂTON: Susan is an present-day archaeologist. She specializes in Aurignacian artifacts, and is currently writing an encyclopedia article on *bâtons de commandement*, mysterious pierced rods of antler found at many Aurignacian dig sites. She writes everything she believes about the rods in the article: that they are made of antler; that they are curved; that they were used for religious ceremonies; that they are of Aurignacian make, etc. Quite unknown to her, and her colleagues, *bâtons de commandement* have none of these properties: they are straight and made of the ulnae of hyenas; they were used as spear-throwers; and they were of Périgorian make.

If this story is even consistent, then '*bâton de commandement*' must be a non-descriptive term, for none of the beliefs that archaeologists hold of these rods are true, and thus satisfying such beliefs is not a necessary condition for being a *bâton de commandement*.

Let's turn to *etiological kinds*. Here I'm thinking of, for instance, footprints, sewage, exhaust, lacerations, stains, photographs, and stigmata. Qualitative duplicates of

footprints aren't footprints; and footprints serve no function; no, to be a footprint is to have a certain etiology—it's to be the imprint of a foot. Some objects are of a kind because they are in a state or have participated in an event with a certain etiology. For example, to be a leper is to have certain symptoms caused by leprosy; and to be an epileptic is to have undergone and to be prone to undergo seizures with a specific etiology. There are also kinds for which a particular etiology is necessary, though not sufficient for membership. For example, sunburns differ from tans, but not by what causes them. Similarly, to be snow is partly to be produced in a certain way (shaved ice isn't snow), but it is also to be constituted by H_2O.

Rust is a good case of an etiological kind. A red iron oxide *not* formed by the corrosion of iron or iron alloys, such as a "swamp" iron oxide formed *ex nihilo*, is not a case of rust (or so I intuit; your mileage may vary). To be rust is to be corroded iron. But consider Becher, who took rust to be dephlogisticated iron and suppose again that he and all those before him were subject to optical illusions, and mistook the observable properties of rust. Then either the meaning of 'rust' changed since Becher's time, or it turns out that there is *now* no such thing as rust (just as there is no such thing as phlogiston). But the meaning of 'rust' hasn't changed (we trust that if Becher had lived to see Lavoisier's experiments, he wouldn't have said that nothing was in the extension of 'rust'), and rust itself is here to stay. Of course, many will say that 'rust' is a natural kind term (although how it is like hearts or lightning is beyond me—and though it's like 'water' and 'cesium,' their etiology is irrelevant to their identity), and this is a respectable position. But let's consider a clear case of an etiological kind that is not also a natural kind, by anyone's lights:

> CANALE: The year is 1877 and Mars is in opposition. Giovanni Schiaparelli, the Italian astronomer, takes the opportunity to carefully examine the red planet via telescope. What he observes is a set of roughly straight lines criss-crossing the planet, which he dubs '*canali*.' Several theories are defended in subsequent years by those who take Schiaparelli's observations to be accurate (a matter which is by no means a consensus). Some think that the *canali* are natural formations caused in a certain way, say, by the flow of liquid water along the planet's surface. Others think that they are literally canals dug by agents who inhabited or once inhabited Mars. Quite unknown to them all, the *canali* are really mounds created by giant underground Martian sandworms (and due to unusual conditions in the Martian atmosphere, don't look at all as they appeared to Schiaparelli in his telescope), and so all current theories concerning them are false. But still, the *canali* exist.

(You can actually generate several examples from CANALE by switching the "true" theory at the end with one or another of the false ones. Fun for the whole family.)

Although examples are fun, I'll leave off giving more for other kinds of non-natural kinds. The recipe for an Argument from Error should be clear. For any general term "G," determine where it falls in the taxonomy, and what its individuating characteristic is (social role, telos, interesting effects, etc.). Then consider two skeptical hypotheses: first, that the linguistic community has a false theory, or no theory, about the individuating characteristic of G (that is, they don't know the social role, telos, interesting effects, or whatever, of G); and second, that their primary means of recognizing instances of G are subject to perceptual or testimonial illusions[2]. It's usually not difficult to find a case where the first skeptical hypothesis is *true*, as with Archimedes and gold, Galen and hearts, and Becher and rust. This is because theories are, in general, hard-won. Under the skeptical scenarios, nothing satisfies most, or a weighted most, of the properties in the cluster associated with "G," and thus by (GT1), nothing should satisfy "G" either. But

[2] That is, unreliable testimony that *appears* to be reliable, such as the fact that books on dream-catchers are included in the 'metaphysics' section of the bookstore.

then just test this hypothesis: imagine the community to be disabused of their unsatisfied theories and cured of illusion. Do they now assent to "there are no G's after all"? And the answer will always be: "presumably not." So condition (GT1) is false for general terms, generally.

4. Non-natural kind terms: arguments from ignorance

In this section, I'll continue introducing various kinds of kind terms, and go on to show how to construct Arguments for Ignorance centered around them. The thesis we'll be concerned with here is GT2, that satisfying most or a weighted most of the properties a group of speakers associates with a general term "G" is a *sufficient condition* for being G:

> GT0. To every non-natural kind general term "G", there corresponds a cluster of properties (of individuals), namely the family F of properties φ such that A believes 'all G's are φ'.
>
> GT2. An object x satisfies "G" (as used by A) if x satisfies most or a weighted most of the properties in F.

Structural kind terms express properties which are instantiated when the objects in their extensions stand in certain relations to other objects, or when the parts of objects in their extensions stand in certain relations to one another. The first division of structural kinds subsumes most human *social kinds*, such as marriages, money, art, back-yards, parties, stores, and philosophers; the second division subsumes many *mathematical kinds*, such as triangles, measures (e.g. probability distributions), and successors. It is not hard to find instances of kinds that are only partially structurally defined—for example, both

the constitution and the structure of a wave determines whether it is red or, say, in the key of G.

Mathematical kinds may at first seem to resist anti-descriptivist arguments, and even to provide evidence against them. More or less anyone can produce on demand necessary and sufficient conditions for being a square. Those who can't are clearly incompetent with the term and furthermore it is difficult to see in virtue of what their concept would mean *square*, if their conception of it didn't distinguish squares from, say, triangles. But not all mathematical kinds are like squares. Consider *continuous functions*[3]. Before Bolzano formulated a rigorous ε-δ definition of limits, mathematicians nevertheless used 'continuity' intuitively[4], and this intuitive conception was in part how they recognized Bolzano's definition to be correct. It's implausible to think that they tacitly possessed what Bolzano ultimately produced, but even so, we would be hard pressed to take even the ε-δ conception of continuity as a meaning-determining definition of continuity, given our subsequent recognition of continuity's more general nature as a morphism between topological spaces (the ε-δ definition articulating continuity in the special case of metric spaces). So it is clear that at least in some cases, we don't possess adequate ideas of our mathematical concepts. Unlike 'continuous,' it presumably didn't take hundreds of years to hit on an adequate set of necessary and sufficient conditions for the application of 'square.' This may be an artifact of the obviousness of such conditions. But if even 'square' is amenable to the sort of generalization 'continuous' underwent in topology, then it might not be right to describe our current "definitions" of 'square' as determining its meaning.

---

[3] The example, and much of the content (though not the assertoric force) of the ensuing discussion, I owe to Ben Levinstein.

[4] Or as part of a false theory involving infinitesimals.

I won't give social kind cases, because they look almost exactly like the '*bâtons de commandement*' case in the previous section: imagine a cultural practice of an alien or long-dead civilization. Give it a name. Suppose we have insufficient information concerning it to distinguish it from similar, but distinct practices. If this is possible, satisfying our descriptive information associated with the name is insufficient for instantiating the relevant social kind. Rinse and repeat.

There are general terms expressing properties whose instantiation involves an arbitrary cutoff somehow. All the measure terms are like this: 'day,' 'week,' 'meter,' 'Newton,' etc. Presumably to have a length is to be composed of physical stuff (space-time included) standing in appropriate structural relations; yet to have a length of one meter is to have a length and also meet some arbitrary standard. To be a day is no less arbitrary: 86,400 seconds is no more a natural carving of the world than 72,500 is. Similarly, man has made many non-natural but humanly useful groupings of objects: separating mountains from hills, planets from Plutos, and even the elderly from the youth. All of these divisions ultimately require a cutoff in what is essentially a plenum. I won't use the word 'kind' for *arbitrary-division properties*, but I will assume they are in fact properties. Pavel Davydov (in personal communication) suggested to me the following Twin Earth style case for 'meter.'

> METER: In a galaxy far, far away there is a planet much like our own, which we may call Tiny Earth. In fact, Tiny Earth is exactly like our own planet, except everything is one one-thousandth of its size on Earth. So for example when Pavel on Earth sits at his computer writing his dissertation, Tiny Pavel on Tiny Earth sits at his tiny computer writing his tiny dissertation. The inhabitants of Tiny Earth are, internally, subjective duplicates of the inhabitants of Earth.

So what's "in their heads" is insufficient for establishing the extension of 'meter' (as used by them), for it's the same as what's "in our heads," and our meters are thousand times longer.

We may adopt the term 'motley' (from Segal (2000)) for a general term "applying, roughly speaking, to anything satisfying the core descriptions associated with them" (p. 54). This definition may seem to define 'motley terms' as just 'descriptive general terms,' but this is a mistake. The definition says that motley terms do, as a matter of fact, apply to anything satisfying the core descriptions associated with them; it does not say that the reason why—the metasemantic ground, so to speak—that they have such extensions is that they are descriptive. For instance, most adult speakers of English associate 'H_2O' with 'water,' and if this is the "core description" they associate with 'water,' then 'water' applies to anything satisfying the core description associated with 'water.' That is, 'water' is a motley term, by definition.

Segal suggests 'chronic fatigue syndrome' as a more paradigmatic motley, though this is more of a name of a syndrome than a general term (not Segal's fault—he doesn't limit himself to general terms. It is I who am being picky.) Jerry Fodor has suggested 'poison' and 'grass' to me (p.c.) and has this to say in his review of Hughes' *Kripke: Names Necessity and Identity*, "Water's Water Everywhere":

> What about *water* makes it necessary that water is H_2O? There must be something about water that does because, notice, there are plenty of kinds of stuff for which the corresponding modal claim would be false... Every sample of smog is a sample of CO_2 and god knows what else; but that's only contingently true. Perhaps tomorrow they'll find a way to pollute the air by using XYZ. Then, ceteris paribus (according to my modal intuitions), the right story would be that they've found a new way to make smog, not that they've found a way to make something that seems just like smog but isn't.

I don't think the existence of motleys should be of much concern to the anti-descriptivist. First, I take it that in all these cases—smog, poison, grass—it's a posteriori that they're motleys. Consider for instance some ancient who is under the false impression that every poison contains the same substance, in virtue of which it possesses its toxicity. From an "internal" perspective, how he conceptualizes poison is just like how he conceptualizes gold. Any metasemantic story must be at least partly externalist—it must take account of whether or not poisons in fact have a common underlying essence. The same can be said of anyone who thinks 'grass' is a species term, smog is an element, or chronic fatigue syndrome is a type of lyme disease.

Furthermore, nothing in the case precludes an entirely externalist causal account. In the case of 'gold,' we're disposed to apply the term to objects that lack the superficial characteristics of gold, but share its underlying essence (when we know that they share it). In the case of 'poison,' we have no such dispositions, because there aren't any such counterfactual circumstances, because poison has no underlying essence. So the externalist can capture the semantic difference between the two terms without going "internal," and only considering the dispositions of agents to token concepts, or apply terms, in counterfactual scenarios. Motleys may not provide any evidence that anti-descriptivism is true, but they don't provide any evidence that it's false either.

The recipe for constructing Arguments from Ignorance for a non-natural kind term "G" is as follows. First, identify the kind's individuating characteristic. Then, consider the skeptical (or actual) scenario where the linguistic community has a false or incomplete theory of the individuating characteristic of G's (a false theory of the etiology of rust, or the nature of eclipses, for example). This rules out the possibility that speakers

possess a theoretical criterion satisfaction of which is sufficient for G-hood. Next, suppose that members of the linguistic community cannot as a point of fact tell the difference between G's and non-G's, as Putnam cannot tell the difference between elms and beeches. This rules out the possibility that speakers possess a practical criterion satisfaction of which is sufficient for G-hood. If such suppositions are in general consistent, it follows that condition (GT2) must be false for "G": satisfying all of the properties speakers associate with "G" simply will not guarantee membership in G. And such suppositions are in general consistent, so (GT2) is false for general terms, generally.

5. Tying it all together

When Kripke discusses names, he lists six descriptivist theses, and argues against each of them separately (except the first, the analogue of our (NK0), which he takes to be definitional). I've never really understood that aspect of *Naming and Necessity*; arguments refuting some of the theses immediately refute other theses. Perhaps historically, hammering it home was necessary. But I suspect no-one wants me hammering anything, so I'll just point out how the suppressed descriptivist theses for general terms are refuted by the foregoing (should the foregoing constitute an actual refutation, of course).

First, there's the thesis that (GT1) and (GT2) are a priori—that not only is it the case that the extension of a non-natural kind general term coincides with the extension of "most or a weighted most of the properties speakers associate with that term," but also that the relevant speakers know this fact a priori. Some versions of descriptivism (the

"community-wide" versions of last chapter) deny this thesis, but some descriptivists, particularly descriptivists motivated to rationalize the a priori methodology of philosophy, would wish to uphold it. However, if the preceding argumentation is correct, and (GT1) and (GT2) are *false*, it follows immediately that they are not known a priori, for nothing false is known, a priori or otherwise. Second, there's the thesis that (GT1) and (GT2) are necessary. Again, what's false at the actual world is of necessity not necessary, because necessity is truth at every possible world, and the actual world is a possible world.

That sounded pretty snappy, but some concessions are in order. What I've shown is not that (GT1) and (GT2) are false for all functional kind terms or all etiological kind terms or all structural kind terms or all arbitrary division terms or all motley terms. Instead, if I've shown anything at all, it's that (GT1) and (GT2) are false for *some* members of each of these groups. And I can't do better, and I shouldn't be expected to. After all, I know that gold has atomic number 79. *Ipso facto*, (NK1) and (NK2) are true for speakers = {me}, "K" = 'gold,' F = {'has atomic number 79'}, and weights = {<1.0, 'has atomic number 79'>}. Furthermore, given these parameters, (NK1) and (NK2) are *necessarily* true. Likewise, I know the function of hearts, the etiology of footprints, the topological definition of 'continuous function,' and the equation for Planck momentum. Setting the appropriate parameters in (GT1) and (GT2) will result in true—indeed, necessarily true—generalizations involving each of the kinds of non-natural kind terms we've been considering.

Any term for which we can produce a metaphysically necessary, necessary and sufficient condition will *look* descriptive, but this no more proves that it *is* descriptive

than the fact that I know that water is H_2O proves that 'water' is descriptive. But once it is established that *some* natural kind terms are non-descriptive, there doesn't seem to be any reason to hold that other natural kind terms have a fundamentally distinct metasemantics. Putnam can't tell the difference between a beech and an elm, but he can tell the difference between a conifer and a flowering plant. Yet this hardly seems to warrant treating 'conifer' in a different fashion from 'beech.' The difference here is epistemic, not semantic. This sort of reasoning seems to defeat internalist versions of descriptivism, but it is clear how to extend it to community-wide versions too. Our linguistic forebears were ignorant of the hidden natures of natural kinds, and often in error when they ventured to form theories. In such cases, no-one in the community possessed individuating criteria for, say, being gold.

I admit that it is more difficult to imagine cases of non-natural kind terms where everyone in the community is similarly ignorant, but I submit that BÂTON, CANALE, and METER, for instance, are such cases. In BÂTON, no-one knows what *bâtons de commandement* are for; in CANALE, no-one knows what causes the *canali*. You cannot simply look at an artifact, or a formation, and know what its function or cause is any more than you can look at a substance and know what its internal structure is. But this does not prohibit you from having concepts or words that apply to such artifacts or formations *qua* artifacts or *qua* formations any more than not knowing the chemical nature of gold prevents you from having a chemical kind term for it. Even the lack of a hidden essence is itself hidden: you can't look at smog or grass or poison or Coca-Cola for that matter and be able to tell that these *aren't* natural kinds. Science isn't just the process of discovering where essences are, it's also a process of discovering where they

aren't. Descriptivism would have us know at the outset, completely and fully, what it is we're talking about. But we are too ignorant and too subject to error for this to be prudent in any case.

Let's grant then that in addition to natural kind terms, several other kinds of general terms are non-descriptive: functional kind terms, etiological kind terms, structural kind terms, terms for arbitrary-division properties, and motley terms. It's still as yet an open question whether these are all the categories of general terms, or whether I've cherry-picked a subset that make my arguments look good.

There's some reason to think that the kinds we've discussed constitute the main categories under which humans tend to classify things. Aristotle thought that there were four sorts of accountings that one could give of objects. For each object, one may say what it is made of (its material cause, as in the substance kind terms), what brought it about (its efficient cause, as in the etiological kind terms), what it is for (its final cause, as in the functional kind terms), and what the structural relations among its parts are (its formal cause, as in the structural kind terms). Now, Aristotle is likely not right in thinking that one can give all four accounts for every object: most objects lack purposes, many lack causes (think radiation from a decaying nucleus), and if absences are things (think singularities), some lack material constitutions. Furthermore, we must also take account of terms that express arbitrary-division properties and motleys, which don't tend to be characterized solely by their material, efficient, final, or formal causes. But I doubt there are many categories of general terms besides those we have considered[5].

[5] I can think of two: since etiological kinds are characterized in terms of their cause, one might suspect that there are kinds characterized in terms of their effects. For example, if the consequentialists are right, '(morally) good' might be such a term. But I don't think I need to give cases to show that we have been ignorant or in error with respect to what's morally good. The other category I can think of is eventive or

6. Objection 1: Descriptivism as a precondition of Kripke-Putnam intuitions

Didn't Archimedes possess a description intensionally equivalent to 'gold'? Didn't Galen possess a description intensionally equivalent to 'heart'? And Becher a description intensionally equivalent to 'rust'? Archimedes thought that gold is the substance shared by *these samples*; Galen thought that hearts were things with the function of *these samples*; and Becher thought that rust was the product of iron exposed to water and air. Is descriptivism then vindicated?

Some have even held that a descriptivism of this sort *must* be vindicated. The arguments adduced by Kripke and Putnam, and similar ones adduced by me, above, all rest on our *intuitions* concerning the truths of certain sentences given certain hypothetical scenarios. For example, an informant is asked to imagine that everything she and other speakers in the community have called 'water' has a certain underlying nature, Q. But everyone in the community is unaware of this and believe firmly what they express as 'Water is X,' for some X that is, say, anti-correlated with Q. She is then asked: is it true, false, or undetermined whether water is Q? If she answers 'true', we take this as evidence that the beliefs of speakers in the community do not determine the meaning of 'water.'

But we may well wonder *why* the informant answered 'true.' A computational model of judging the truth-value of a sentence in some context, say, our story as told by the linguist to the informant, might say something like: to the context C (I want this to be some sort of mental representation of the story) the informant adds her background

---

stative terms like 'eclipse,' 'hurricane,' 'snap (of the fingers),' or 'sleep,' where the nature of the relevant property is a state or event of some sort. And we wouldn't need to dredge history to find cases where people were in ignorance or error with respect to the nature of eclipses, say, or sleep.

information B and definitions D. If the linguist wants to know whether *P in C*, the informant answers 'yes' if she (the informant) deduces P from C & B & D, 'no' if she deduces ~P from C & B & D, and 'undetermined' if she deduces neither P nor ~P from the premises[6]. If our informant answers 'true', as above, then either her background beliefs or her definitions (in conjunction with C) must entail that water is Q—that is, she must already know that water = whatever has the underlying nature of the stuff called 'water'. If this is not contingent background information, then to even have Kripke-Putnam intuitions in the first place requires speakers to possess a priori necessary and sufficient conditions criterial for the application of natural kind terms.

This is how I read Donnellan, for instance, when he says, "That in some important sense the rules are followed by us [read: definitions are represented by us or embodied in our behavior] in the case of terms for natural kinds is perhaps attested to by the force of the intuitions about what we would say in this or that imagined case" (p. 158).

I shall argue, however, that descriptivism is not vindicated. I suspect that Galen *did* conceptualize hearts as organs with the function of those he studied, and I do not deny that 'heart' is intensionally equivalent with that description. '*καρδιά*' in Galen's mouth likely did satisfy conditions (NK0)-(NK2). But this was a coincidence. Even theories about the essences of the real essences of our terms are hard-won, and they too may be opaque to us.

Putnam responds to this objection as follows (see The Meaning of 'Meaning' pp. 242-5). We might (in the epistemic sense of the word) discover that pencils are organisms. The relevant, though far-fetched example is one in which the objects we are

[6] Notice that the informant will answer 'undetermined' when the computations are too difficult, even if C & B & D *entails* P.

disposed to call ‘pencil’ are living beings that feed, reproduce, age, and die—though mostly out of sight. Putnam’s intuition, one that I share, is that this case is correctly described as one in which pencils are organisms. If this is so, then ‘pencil’ cannot be synonymous with ‘objects with the function F’ (for an adequate description of pencil-function F), for organisms haven’t any function, much less F[7].

Although in the case of gold, hearts, and pencils we *know* at the outset true theories of the nature of their real essences—that is, we know that gold is a substance, and that hearts and pencils have functions—this is not true in the general case. The fist thing to consider is certain skeptical scenarios where we lack such knowledge. For instance, gold might (epistemically) have been a motley. Given all that Archimedes knew about gold, it still could have turned out that there was no underlying substance or hidden essence of gold—that some gold was cesium, other gold mercury, still other gold helium, etc. So even though Archimedes in fact knew that gold was a substance (I don’t assume you need to rule out skeptical scenarios to have knowledge), he didn’t know this a priori, since qualitative duplicates of Archimedes in other worlds (worlds where what plays the gold-role is a motley) don’t have this knowledge. Similarly, though this is somewhat far-fetched, hearts might have turned out to be evolutionary spandrels, and as such be totally bereft of a function. Yet upon discovering such, we wouldn’t suppose that there weren’t any hearts. Even when we possess knowledge concerning the true, underlying natures of various kinds, it does not infect the semantics for these terms, and this is for the best. For, if we stipulated that pencils had to be artifacts, that gold had to be a substance, or that hearts had to have functions, skeptical scenarios like that described by Putnam would

[7] I haven’t entirely convinced myself of the truth of this claim. If I used a clam as a paperweight, or a monkey as a butler, would they then be *for* weighing down my papers and fixing my drinks, respectively? If so, then Putnam’s argument as stated won’t do. See the next paragraph for more compelling examples.

threaten to undermine even our mundane knowledge of pencils, such as that I possess ten of them.

And these cases aren't mere skeptical scenarios. Suppose you arrive on some alien world and discover a bunch of strange tree-ish objects, which you see regularly enough to call 'strees.' Are they organisms? Natural formations? Alien artifacts? You might not know. So you are not in a position to stipulate 'strees are whatever has the function of these things' or 'strees are whatever has the underlying physical nature as these things,' or whatever. Suppose that in actuality they're really big fundamental particles—treetrinos. And suppose further you had the following linguistic disposition: upon discovering that strees are in fact treetrinos, you would be disposed to assent to claims like "there could be things that perfectly satisfy all the things I knew about strees when I introduced the term, but were nevertheless not strees, in virtue of not being treetrinos." I rather suspect that most people in that same situation would have the same disposition. It thus can't be that these intuitions are driven by a priori semantic stipulations or meaning postulates.

This should resolve our initial puzzle, concerning the force of our intuitions in the Kripke-Putnam cases. Linguistic informants who answer "water is indeed Q" when asked to suppose (perhaps counterpossibly) that the underlying nature of the stuff called 'water' in their community is Q must, I submit, represent to themselves that water = the stuff with the underlying nature of the stuff called 'water' in their community. But this cannot be some piece of a priori, definitional knowledge possessed by these speakers, because the selfsame speakers will intuit that 'water' is not a substance term, given suitable distinct contexts, where what explains the common features exhibited by the stuff called

‘water’ is something other than the constitution of the stuff. Rather, informants employ a posteriori (and perhaps false) background assumptions, including theories about the natures of properties expressed by their predicates, when their intuitions are elicited. Descriptivism is falsified, rather than vindicated, by the method of cases.

7. Objection 2: competence

The second objection I’d like to consider is that some general terms are *obviously* ‘descriptive’ in some sense or another, because speakers who don’t associate the appropriate descriptions with those terms are incompetent or not fully competent with the term. Consider the following speech:

> “Poor Fred, he doesn’t know what ‘bachelor’ means. Yesterday I heard him wonder aloud: ‘Are potatoes bachelors? Perhaps only Russets?’ He clearly hasn’t mastered the term, for he doesn’t know that bachelors are unmarried men. He is incompetent vis à vis the term ‘bachelor.’”

This speech, if it is correct in point of fact about Fred’s behavior, is correct in point of evaluation of Fred’s mastery. Fred *hasn’t* mastered the term; he *isn’t* competent; and he *doesn’t* know what ‘bachelor’ means. This is all well and good, but I don’t see what it has to do with semantics. Let me explain.

In virtue of what does a word mean what it means, and not something else, or nothing altogether? This is a difficult question; I don’t know the answer; and I certainly don’t want to prejudge the case in favor of one or another theory. What I do want to point out is that not all plausible answers to the question validate the inference from “Speaker S is not competent with term T” to “Term T doesn’t mean in S’s mouth what it does in

everyone else's." If this is so, we should be very hesitant to make or accept such inferences.

Consider the answer to this question that Fodor gives: the asymmetric dependence theory (1987, 1990). According to Fodor, a term T means some property P in the mouth of speaker S, if P's cause S to token T's, and non-P's that cause S to token T's do so only because P's do (and not vice versa). Now consider George. George does not know that bachelors aren't potatoes. But he does have a new iPhone with the Google Bachelor application. Google Bachelor allows you to type in a name or description of any object and returns the answer to the question of whether that object is or is not a bachelor. Now, George is clearly not competent with the term 'bachelor'; but his uses of the term do depend lawfully on the bachelorhood of bachelors, and his misapplications of the term depend on the first regularity (but not vice versa). So at least on the asymmetric dependence theory, competence is no guide to semantic value.

It would be unwise, I think, to say "so much the worse for the asymmetric dependence theory," and leave it at that. First, many other theories entail this same result. On Dretske's (1988) view, roughly characterized, the 'bachelor' concept means what it does because: it has been promoted to control-duty over an agent's behavior because it carries information about bachelors. Surely these conditions may be satisfied by an agent who does not know that bachelors are unmarried men. For instance, 'bachelor' may be successfully recruited to control appropriate bachelor-directed behavior because of its informational connection to bachelors via the agent's use of Google Bachelor.

The second reason is that there are general theoretical reasons for divorcing competence judgments from our data set for semantic theorizing. Judgments of

competence seem to track a word-user's ability to convey semantic proficiency to others. George is incompetent precisely because he cannot tell *you* how to apply the term 'bachelor.' He can tell you how to acquire the ability (go get an iPhone and download Google Bachelor), but he cannot give you the ability through instruction. Putnam indeed thought that lack of the tiger stereotype rendered speakers incompetent in their use of 'tiger' for just this reason [cite], namely, that they hadn't any representation that would allow them to recognize and respond appropriately to tigers. But all this was so, for Putnam, even if the causal and social connections those speakers bore to tigers made their words mean *tiger*.

It is not, however, clear that we want to require speakers to have descriptive or recognitional capacities to mean, say, *bachelor* by 'bachelor' (as opposed to being competent in the use of the word). If George wonders aloud 'most things I've discovered to be bachelors are unmarried men… are bachelors, perhaps, all and only the unmarried men?' it would be strange to say that he must be speaking nonsense, or not fully grasping a thought, just because he doesn't already know the answer to his speculation. Strange I say, but not impossible.

Second, it's not clear how it could be true that anyone was incompetent with 'bachelor' if on the one hand 'bachelor' were descriptive, and had its meaning determined by the beliefs of its users and on the other hand, competence with a term required possessing the relevant meaning-determining beliefs for that term. Suppose for a moment that George doesn't merely wonder whether russets are bachelors, but rather believes with great conviction that they are. Then, since George associates 'is a russet' with 'bachelor,' giving it great weight, if 'bachelor' as used by George applies to

anything, it applies to russets. Thus, George possesses the relevant meaning-determining descriptions for ‘bachelor.’ Thus, George is competent. This isn’t, by the way, an argument that descriptivism is false (see the previous sections for that), it’s an argument that descriptivism can’t both be a metasemantic theory and at the same time be a competence theory.

Finally, I’m not even sure it’s appropriate to deny competence to the majority of speakers in my examples above. Consider Susan, the archaeologist in BÂTON. She is the world’s leading expert on *bâtons de commandement*. Her views are more justified than anyone else’s. Furthermore, no-one has a better theory of their function than she. If she’s not competent, no-one is; and if lack of competence with a term entails an abnormal, defective, or empty semantic value, then being as no-one is competent with ‘*bâtons de commandement*,’ it is meaningless drivel. Go tell that to the archaeologists at your university, that if they happen to be mistaken about their theories, then their work is nonsense. I don’t suspect it will go over too well.

To summarize: I am not here claiming *in this section* to have demonstrated that ‘bachelor’ is non-descriptive, nor have I produced any arguments to that effect *in this section*. What I am claiming is that the inference from our intuitions about when a speaker lacks competence with a term, to the claim that the semantic value of the term in the speaker’s mouth is empty, partial, or abnormal, is an inference of dubious merit. At the very least, it is an inference that should take a back-seat to inferences involving those of our intuitions that directly concern the semantic values of terms in speakers mouths. And it is these latter data points that I presented as my main arguments here.

8. Conclusions

One objection[8] not considered in the preceding sections is the charge of *profligate properties*. This charge alleges that properties are sparse, and furthermore, that the only ones that exist are the natural properties (natural kinds). It would then seem to follow that the extensions of non-natural kind terms must be determined by a cluster of natural properties, rather than themselves directly referring to non-natural properties (which, on the view in question, don't exist).

I am no metaphysician, so I have nothing to say concerning whether such a view is or might be true. What I will say is that if it's true, it's nevertheless irrelevant to the thesis I wish to advance. What's wrong with descriptivism for non-natural kind terms, as I see it, is that it is internalist, that the extension of a general term at a world is determined by a cluster of properties *speakers* associate with the term. We have seen this already to be false. If the one true metaphysics requires that a cluster of properties determine the extension of non-natural kind terms, but allows that this cluster be associated with the term by factors external to speakers, then I have nothing to say against that metaphysics, and I wish it the best of luck.

My main conclusion is that there is nothing distinctive about the semantics of natural kind terms. Kripke and Putnam could have conducted all their semantical investigations into externalism using non-natural kind terms. They could have used artifact kind terms like 'hammer' and 'table' and run all of the same arguments, as I do above. The conclusions are just as compelling, and for all the same reasons. The proper lesson to draw from the writings of Kripke and Putnam is not that names and natural kind

[8] Thanks to Will Starr for articulating the objection to me.

terms stand apart in uniquely being non-descriptive, but that *all* terms great and small are non-descriptive.

It appears that now we have strong inductive evidence that *all* words are non-descriptive. Establishing this claim on firmer ground is an important task for philosophical semantics. For standard descriptivist responses to Kripke and Putnam look far less appealing when applied wholesale to every word. For example if we extend the account that 'Cicero' is shorthand for the description 'that which people call 'Cicero'' and 'water' means 'that which people call 'water'' to all words, we get a fairly vacuous lexicon:

> 'Aardvark' means: that which people call 'aardvark'
> 'Abacus' means: that which people call 'abacus'
> …
> 'Zygote' means: that which people call 'zygote'

The two most important upshots of the view, however, are those touched upon in the introduction. First, if descriptivism is false, not just for proper names and natural kind terms, but for all semantically primitive expressions (morphemes), then metasemantics (the discipline that answers the question: "In virtue of what does a primitive expression have the semantic value it does, rather than some other semantic value, or none at all?") can finally be uniform: we have not two stories for how morphemes get their meanings, but one. This will presumably be a causal-historical story, though not a baptismal one, given the results of Section 6.

Second, wholesale anti-descriptivism allows us to resist an untoward semantic anatomism. If non-natural kind terms had their meanings constitutively determined by those descriptions speakers (or communities of speakers, or groups of experts in communities of speakers) actually associated with them, then two speakers could never

possess *different* theories of some artifact, phenotype, natural formation, social practice, or mathematical object: they would have to possess the same theory, or be talking past one another. However, if no term derives its meaning from descriptive information speakers actually possess, but instead all rely on a causal connection between concept and property, then speakers may go about constructing theories of important classes of objects outside themselves without substantive advance knowledge of what their theories are to be theories of (that is, allowing for natural human ignorance and susceptibility to illusion).

That uniform causal theories offer a general solution to the paradox of inquiry may explain why they are true. As children, we are thrown into a world of individuals, substances, animals, artifacts, and customs with little knowledge of their real essences. Theories are hard-won, and the most diligent inquiring minds among us must rest content doomed to inadequacy and error on all fronts. We cannot say in advance what we will find, even in the most abstract way; and further it would be foolish to select some subset of what we *might* find (by stipulating definitions before our inquiries have been conducted) and build only theories concerning that. No, we must take what there is, what we stand connected to, what our concepts represent will we nill we, never being so rash as to claim victory (by stipulating definitions before our inquiries have been completed), ever aware of what we don't know and may never know.

Chapter 2

Rigidity for the Common Noun

I argued in the last chapter that all general terms (and perhaps all adjectives… and *perhaps* all words) were non-descriptive. It remains however to be seen whether they are all rigid, and further whether they all appear in generalizations that are both necessarily true and a posteriori.

The question of whether general terms are rigid is vexed, in large part because there is widespread disagreement over and confusion concerning what rigidity for general terms amounts to. Kripke himself seems to have vacillated on the question, as Salmon reports:

> Responding to my comments during the discussion of Soames's presentation at the 1996 *Universidad Nacional Autónoma de Mexico* conference (see footnotes 3, 11 [in Salmon]), Kripke said that this [i.e. Salmon's] proposed interpretation of N&N on general-term rigidity is basically correct. Soames reports that in November 1997, when he presented what is essentially the same interpretation proposed in the book [i.e. Salmon's] with Kripke in attendance, Kripke this time expressed sympathy with Soames's assessment that there is no notion of rigidity for general terms relevantly analogous to singular-term rigidity. [2005 p. 366 n. 22]

Kripke defined the notion of rigidity for singular terms as follows: "Let's call something a *rigid designator* if in every possible world it designates the same object" (*Naming and Necessity*, p. 48). But he never provided a definition of rigidity for general terms[9], and we can't simply apply the definition as-is. Natural kind predicates in the main are not singular terms, that is, their semantic values aren't objects or individuals. Soames (2002,

[9] Soames makes this point on p. 245 of *Beyond Rigidity*. At that time, he says, the point was "not widely appreciated," though the subsequent literature seems to have heard him. For an alternate view, Salmon (2005b, p. 120 n. 6) says "I believe Kripke intended his definition of rigidity to apply to general as well as singular terms."

p. 264) points out, for instance, that although one might make a case for the mereological fusion of all water being the referent of 'water', it is patently implausible that there is an *individual* that is the referent of 'cat' or 'hot.' Furthermore, since other worlds have more or less water than ours, 'water' comes out non-rigid on this proposal. It thus seems that we require a separate account of rigidity for general terms.

Several considerations speak to the importance of answering these questions. We can see this by considering why Kripke's discovery that *names* were rigid designators is philosophically significant. First, only if variables are rigid designators can we hope to validate Leibniz's Law for quantified modal discourse, i.e. $(x)(y)(x = y \rightarrow (\varphi \leftrightarrow \psi))$, where the (possibly) quantified, (possibly) modal formulae φ and ψ differ at most in that free occurrences of x in φ are replaced by occurrences of y in ψ; and only if all singular terms are rigid designators does every substitution instance of Leibniz's Law, e.g. $a = b \rightarrow (\varphi \leftrightarrow \psi)$, turn out valid. Since intuitively, all identical objects have identical modal profiles, such validation is highly desirable. A further consequence is that a modalized Leibniz's Law (plus Rule N) entails the necessity of identities: $(x)(y)(x = y \rightarrow \Box\, x = y)$. Thus the rigidity of names accounts both for our intuition that identical objects have identical modal profiles and for our intuition that if two things are one and the same, they could not have been distinct. In addition to this point, there is a far more straightforward way in which a rigidified treatment of terms simplifies our modal logic, as terms need not be assigned semantic values (individuals) relative to a world parameter, but may be assigned such values *simpliciter*.

The philosophical significance of Kripke's account would carry over quite directly in the case of general terms. For Leibniz's Law for general terms seems as

transparently true as Leibniz's Law for singular terms: numerically identical properties have numerically identical modal profiles, or[10] $(X)(Y)(X \equiv Y \rightarrow (\Phi \leftrightarrow \Psi))$, where X and Y range over properties, and Φ and Ψ differ at most in that open occurrences of X in Φ are replaced with Y in Ψ. As we shall see, on some accounts of rigidity for general terms, if A and B are rigid general terms, then Leibniz's Law + N entails $A \equiv B \rightarrow \Box A \equiv B$, e.g. if water is H_2O, then water is necessarily H_2O. But not on all accounts.

Resolving this issue is especially important in light of the use Kripke makes of this principle. For in the third lecture of *Naming and Necessity*, Kripke uses the claim that 'pain' and 'C-fiber stimulation' are rigid designators to argue that if pain is actually identical to C-fiber stimulation, then it must be necessarily identical to C-fiber stimulation. On at least one popular account of rigidity for general terms (Rigid Application, considered below in Section 2), this argument form is invalid. Thus, if we're to know how to evaluate Kripke's argument, we must first know what rigidity for general terms amounts to.

Finally, although names seem so different from common nouns in English, in that they don't typically co-occur with determiners, aren't typically modified by adjectives, etc., there's some reason to suppose that they might nevertheless be general terms (Burge, 1973). After all, in some languages, such as Greek, names do co-occur with determiners. Furthermore, even in English they sometimes co-occur with determiners ("Every Fred in the class," "The London of my youth") and sometimes co-occur with adjectives ("Poor, sad, unfortunate George"). I don't want to endorse the line that names *are* general terms,

[10] I here use the symbol '≡' for property-identity. Although this principle seems suspiciously similar to Frege's Basic Law V, it is in fact innocuous; see Section 3 for discussion.

but I will say that to the extent that it's a plausible thesis, we have that much more reason to want an account of rigidity for general terms.

1. Three accounts of rigidity

The striking results we want to capture with our notion of rigidity for general terms are that certain theoretical statements like 'water is H_2O' and 'tigers are animals' are necessary, if true. This suggests two desiderata on an account of rigidity. First, it must be the case that these statements fall under the purview of the account: that 'water,' 'H_2O,' 'tiger' and 'animal' all turn out to be rigid. Second, it must be the case that the account entails that the rigidity of these expressions explains why, for example, it's necessary that tigers are animals. We can summarize this by saying that such an account must be both ADEQUATE and EXPLANATORY:

> ADEQUATE: The account should count as rigid all or most of the paradigm cases of rigid predicates—substance kind terms, secondary quality predicates, natural phenomenon terms, and biological species terms.
>
> EXPLANATORY: The account should explain the necessity of theoretical identities (like 'water is H_2O') and other a posteriori theoretical statements (like 'tigers are animals').

A third desideratum suggests itself when we consider Kripke's definition of singular term rigidity. His definition, when the object designated is made explicit, takes the form of a "semantic persistence claim": a singular term S is rigid iff *if* S designates some object O relative to the actual world, *then* S designates O relative to every other world[11]. That is, what it is for a singular term to be rigid is, first, for it to bear a certain semantic relation

[11] Here I assume that a term can designate objects relative to worlds in which those objects don't exist. Nothing I say shall hinge on this assumption.

(in this case, designation) to a thing (in this case, an object) in the actual world, and second for this semantic relation to persist across worlds: for it to bear that selfsame relation to that selfsame thing relative to non-actual worlds.

If a concept in our explanatory repertoire is to deserve the name 'rigidity,' it should naturally extend Kripke's notion of singular term rigidity. Or rather, rigidity for general terms should be some form of semantic persistence, as it is for singular terms[12]:

DESERVING: The account should make rigidity a form of semantic persistence.

This desideratum is powerful, because it entails that there are exactly as many candidates for general term rigidity as there are semantic relations R that general terms can bear to things. A general term is a rigid R-er iff *if* it bears semantic relation R to thing T relative to the actual world, *then* it bears R to T relative to every other world. There are three semantic relations[13] that general terms can bear to things. First, a general term *expresses* a property; second, a predicate *extends* an extension—the extension extended by a predicate (relative to a world) is the set of things that possess the property the predicate expresses (in that world)[14]; finally, a predicate *applies to* objects—in particular, it applies

[12] The desiderata that a theory of rigidity for general terms must be ADEQUATE, EXPLANATORY, and DESERVING are roughly analogous to Soames' desiderata (ii), (iii), and (i), respectively on p. 263 of his (2002). They appear here slightly modified, slightly differently motivated, and with more colorful names.

[13] It's not easy to say what I mean by "semantic" here. I want to include the relation that an expression bears to its literal contribution to the proposition expressed by larger expressions that include it, but I also want to include other relations, like the relation a definite description bears to its denotation, when it denotes, even if what the description contributes to a proposition is a second-level function, and not an individual. Perhaps "intentional relation" would be closer to the mark.

[14] Lewis (2001) p. 50 takes extensions to include all the individuals that satisfy the predicate in any world. This doesn't make sense unless individuals are world-bound (there's a function from individuals to the world they inhabit)—for otherwise, individuals might both be in and not be in one and the same extension. But with such a function, Lewis extensions are inter-definable with intensions, and I therefore count them as such; see Section 3.

to each object in its extension[15]. We thus arrive at three potential candidates for general term rigidity[16]:

> [Rigid Expression] A general term G is a rigid expresser iff if it expresses the property P relative to the actual world, then it expresses P relative to all worlds.
>
> [Rigid Extension] A general term G is a rigid extender iff if it extends the extension E relative to the actual world, then it extends E relative to all worlds.
>
> [Rigid Application] A general term G is a rigid applier iff if it applies to the object O relative to the actual world, then it applies to O relative to all worlds.

In what follows, I will use the capitalized titles 'Rigid Expression,' 'Rigid Extension,' and 'Rigid Application' to refer to three theories, namely, the theory that general term rigidity is rigid expression, the theory that it's rigid extension, and the theory that it's rigid application, respectively. I will use lower case 'rigid expression,' 'rigid expresser,' etc. to refer to the properties defined above. Notice that even if Rigid Extension, say, is false, some terms may well still be rigid extenders.

Rigid Extension is initially very unintuitive. Extensions are useful for formal modeling, but there is little reason to suppose that they are "real" semantic contents. For, as Sullivan (2007) points out, "the semantics of 'tiger' does not change every time a new tiger is born" (p. 4b). Furthermore extensions don't do many of the things theorists would like to do with semantic values: we don't stand in causal or informational relations to them, and being by definition extensionally individuated, they aren't fine-grained enough to play any serious role in cognitive psychology.

---

[15] I actually take it that there is yet a fourth semantic relation predicates bear toward things, namely the relation they bear toward their intensions. The fact that this semantic relation is persistent turns out to be a trivial consequence of how intensions are characterized; see Section 3 for more discussion.

[16] Note: By 'relative to a world' I mean *relative to a world of evaluation*, rather than *relative to a world parameter in the context of utterance*.

Moreover, Rigid Extension is neither ADEQUATE nor EXPLANATORY. ADEQUACY requires that any account of general term rigidity should count all or most of the paradigm cases of rigid predicates as rigid. But ‘water’ is not a rigid extender: there could be more or less water than there actually is, and in such worlds the extension of ‘water’ diverges from its actual extension. Similar remarks apply to the other paradigm cases: there could be more or fewer yellow things; more or less lightning; and more or fewer tigers (this point is made also by, among others, Soames (2002) p. 250 and Devitt (2005) p. 140). In light of this, Rigid Extension is also not EXPLANATORY, in that it doesn’t explain why necessarily, water is H_2O. Certainly, if ‘water’ and ‘H_2O’ were rigid extenders, it would follow that ‘water is H_2O’ expresses a necessary truth, if it expresses an actual truth. But neither ‘water’ nor ‘H_2O’ is a rigid extender, so this is clearly not the reason that necessarily, water is H_2O.

We are then left with two potential candidates for an account of general term rigidity, Rigid Expression and Rigid Application, and these are the two accounts most popular in the literature. The thesis that general term rigidity is rigid application is held by, among others, Cook (1980), Wiggins (1980), Devitt & Sterelny (1999), Devitt (2005), and Gómez-Torrente (2006). The thesis that it is rigid expression is held by, among others, McGinn (1982), Donnellan (1983), LaPorte (2000), Salmon (2005b), Sullivan (2007), López de Sa (2008), Martí, G. & Martínez-Fernández (2010), and me.

2. Rigid Application

Rigid Application has several appealing aspects. The view says that if a general term G applies to an object O relative to the actual world (or any world for that matter), then it applies to O relative to all other worlds. It certainly seems ADEQUATE. Consider a chunk of gold, X. Could there be a world where X exists, but is, say, composed entirely of plutonium? It seems not. Kripke himself defended the much broader thesis that even a table composed of wood could not have existed, without being composed of wood. However strong this intuition is, it's certainly no stronger than the intuition that pure instances of chemical substances could not have existed, without being composed of the chemical substance of which they are actually composed. So 'gold' seems to be a rigid applier. Similar reasoning seems to apply to many of the other paradigm instances of natural kind terms—e.g. biological species terms and natural phenomenon terms. If Y is a tiger, it beggars belief to suppose that Y might instead have been an antelope.

Additionally, although we have refrained from making it a desideratum on any account of general term rigidity that it count all or most non-paradigm cases of rigid general terms as non-rigid, nevertheless it might be thought that doing so is a virtue of any account of general term rigidity[17]. And Rigid Application delivers nicely. Consider certain terms that aren't paradigm cases of rigid general terms: 'bachelor' and 'table.' I myself am a bachelor, but had things gone otherwise, I would not have been. I sit here at this table; in some other world, a giant uses this very table as a hat. There, it is not a table, and was never intended to be so: it has always been a hat. So neither 'bachelor' nor 'table' is rigid, if Rigid Application is the correct account of general term rigidity. And this is an appealing result.

[17] Though see Section 3 for reasons why this might not be virtuous.

Finally, Scott Soames has pointed out that there is some measure of textual support for the claim that Kripke entertained the notion that general term rigidity was rigid application. For Kripke says "'pain' is a rigid designator of the type, or phenomenon, it designates: if something is a pain, it is essentially so" (Kripke, p. 148; quoted in Soames, p. 252). The reader is referred to Soames' thorough discussion of the nuances of the passage and its context (pp. 252-254); here, I shall just take it as straightforward evidence that Kripke, at least at times, took rigidity for general terms to be rigid application. And although Kripke's beliefs should in no sense be taken as gospel here, they should not be out-and-out ignored either.

In this section, however, I shall argue that the support for Rigid Application vanishes upon closer inspection. In particular, I shall argue that the view isn't ADEQUATE because according to it, *almost no general term is a rigid applier*; and that it isn't EXPLANATORY because as defenders of the view like Devitt (2005) admit, it fails to explain why such statements as 'All and only water is H_2O' are necessary, if true. As such, Rigid Application is worse off than Rigid Extension, and we have already abandoned the latter as a hopeless non-starter.

The worry that Rigid Application isn't ADEQUATE is one that is at large in the literature. Martí (2003), for example, points out certain adjectives that Kripke took to be rigid, such as 'hot' and 'yellow,' are not rigid appliers: "a yellow dress could be dyed and a yellow house could be repainted" (p. 132). The form of the argument is that dresses and houses are objects to which 'yellow' sometimes applies; yet they persist through changes in color; and thus they *would* persist through such changes in color, that is, they *could* exist in worlds where they were not yellow.

Neither Devitt nor Martí see this objection as damning. Devitt (2005) responds to the counterexamples by claiming "it is a mistake to think that the primary task of the rigidity distinction is to distinguish natural kind terms from nominal kind terms. The primary task is to distinguish kind terms that are not covered by a description theory from ones that are" (p. 154). I don't think this reply will help the Rigid Application theorist. First, there are seemingly descriptive terms that are rigid appliers. If any general terms are descriptive, mathematical terms introduced by explicit stipulation are, such as 'prime number' which is defined as a natural number whose only divisors than itself and one. And yet, these seemingly descriptive predicates are invariably rigid appliers: no number is prime in this world and composite in some other. Second, there are non-descriptive expressions that are not rigid appliers. Kripke, for example, spent a great deal of time convincingly arguing that 'red' and 'hot' are not amenable to a descriptive treatment, and I argued much the same for general terms generally in the previous chapter—yet few such terms are candidates for rigid appliers. If the notion of rigidity is supposed to separate the descriptive from the non-descriptive, then rigid application gets 'hot' and 'yellow' wrong, and likely everything else.

Martí proposes a more circumspect retreat: "given the wide variety of terms [that Kripke gives as examples of rigid expressions], it would not be surprising if some of them were not to be in the final cut" (p. 132). The idea then is that someone who endorses Rigid Application can simply bite the bullets of 'hot' and 'yellow', while hoping that there aren't many more bullets. And this is my worry, that every general term is a bullet to bite; or, put more succinctly, that they are none of them rigid appliers (save again for special cases involving logical or mathematical terms).

Among the paradigm examples of rigid adjectives and general terms are ‘hot’ and ‘yellow’; ‘heat’ and ‘light’; ‘gold’ and ‘cesium’; and ‘tiger’ and ‘chimpanzee’. The adjectives have already been dealt with. ‘Heat’ is a curious case. Kripke (1980) suggests that ‘heat is molecular motion’ has the following analysis: “For all bodies x and y, x is hotter than y if and only if x has higher mean molecular kinetic energy than y” (p. 138). Clearly, although this statement is true and necessary, still a pair <x, y> might well satisfy the ‘is hotter than’ relation in one world, and fail to satisfy the same relation in another. By such reasoning, ‘heat’ turns out to not be a rigid applier (for the same point, see Soames (2004) pp. 85-86).

Consider also ‘light’, which Kripke takes to be in the same category as ‘heat’ (natural phenomenon terms). I assume that ‘light’ applies to streams of photons, since all light *is* a stream of photons, and ‘stream of photons’ applies to streams of photons. Streams of photons that are light persist through changes that make them non-light. For example, the so-called relic radiation of the Big Bang (a.k.a. the cosmic microwave background radiation) consists of light that has been redshifted so much it is now microwave radiation, rather than light. If ‘light’ were a rigid applier, such redshifts would be impossible; but this would put Rigid application at odds with physics.

It might well be objected that ‘light,’ as used by physicists, applies to microwaves, radio waves, gamma rays, etc., and thus that my supposed counterexample is really no counterexample at all: light cannot become non-light, no matter how redshifted or blueshifted. Perhaps. However, there is certainly a meaning of ‘light,’ the common meaning, which is more or less “visible light.” The *Compact Oxford English Dictionary* gives us, as its first definition of light: “light, n. the natural agent that

stimulates sight and makes things visible; electromagnetic radiation from about 390 to 740 nm in wavelength." This sense of 'light' seems paradigmatically rigid: "light (in the common sense) is electromagnetic radiation with wavelengths ≈ 390 to 740 nm" is both necessary and a posteriori. This sense seems to have survived the discoveries of physics: indeed, such claims as "microwaves cook with light" and "the Incredible Hulk was the result of an experiment with light gone awry" strike me as patently untrue. So Rigid Application might be able to handle 'light' in the scientific sense, but it makes the wrong prediction regarding 'light' in the common sense.

The case of the chemical elements is somewhat similar. Neutrons are composed of two down quarks and one up quark, protons of one down quark and two up ones. Neutrons bound in unstable nuclei may undergo beta decay, where one (and only one) of their down quarks changes flavor to (i.e. becomes) an up quark through the emission of a W^- boson, thus transforming them into protons. For example, a cesium 137 atom (55 protons + 82 neutrons, for a total of 137 nucleons—at 3 quarks a piece for 411 quarks in total) may release a W^- boson and decay to an atom of barium. To be clear, W^- bosons are force-carriers—all they do is mediate the flavor change. What happens in the interaction is that *just one* of the cesium 137's 411 quarks goes from flavor = down to flavor = up. This is sufficient to convert the atom from cesium 137 to barium 137.

Why is this relevant? Well, if 'cesium' is taken to be a rigid applier, then any object (in this case, atom) to which 'cesium' is correctly applied cannot persist through beta decay, because that would result in 'cesium' *not* applying to it. The thesis that chemical kind terms are rigid appliers turns out to be equivalent to a particular thesis about the persistence conditions of atomic nuclei, just as the claim that 'light' is a rigid

applier is equivalent to a particular thesis about the persistence conditions of photons. And I, at least, find it natural to suppose that atoms can persist through the flavor change of one of their quarks. At least, this is more natural to me than supposing cesium 137 atoms regularly pop out of existence, only to be replaced by barium 137 atoms.

The final class of paradigmatic rigid general terms are the names of species and higher taxa. I can do no better here than briefly summarize the points LaPorte (1997) makes on this score[18]. LaPorte reports that "theor[ies] about what determines the boundaries of species… tend to fall into three camps: the interbreeding approach, the ecological approach, and the cladistic approach" (p. 101). He argues that on none of these standard approaches does an organism belong to the species it does essentially. Those who take the interbreeding approach hold that species are reproductively isolated, interbreeding populations of individuals (p. 101). But whether two populations are reproductively isolated, and whether all their past members were reproductively isolated, is clearly in many cases a contingent feature (p. 102). The ecological approach yields a similar conclusion: on this approach, species are populations that occupy their own unique ecological niche (p. 101). This too clearly makes one's species a contingent matter, for those members descended from the ancestral population who came to occupy that niche might never have done so, for instance if the niche were never to have existed (pp. 101-102). I leave aside the case of clades (pp. 102-104), but suffice to say they're no boon to a Rigid Application theorist.

Devitt responds (2005, p. 147) to LaPorte's objections, though to my mind not convincingly. First, Devitt claims we have reason to suppose "there is an intrinsic

[18] I limit myself here to species, and do not summarize LaPorte's discussion of higher taxa. Instead, I recommend that you read his paper: "Essential Membership." (See the Works Cited.)

component, as well as a relational one, to the essence of a species; in particular…a species has a genetic essence." This reply only works if having the genetic essence of, say, a tiger, is a *sufficient* condition for tigerhood, and this is denied on all the relationalist approaches. For example, on the interbreeding approach, I might well hold that each actual member of *Pongo abelii* (the Sumatran orangutan) essentially has the genes it does, while nevertheless maintaining that each *would be* a member of *Pongo pygmaeus* (the Bornean orangutan), or of some distinct parent species, had the flooding of Sundaland never occurred after the last glaciation, and had all remained in one interbreeding population.

Devitt also urges that "any member [of a species] has [the relational component of a species' essence] essentially if Kripke is right in thinking that an organism's essence is its relation to a certain sperm and ovum, hence to certain parents, hence to a certain family tree" (p. 147). This again avoids the charge. Both the interbreeding approach and the ecological approach allow that one and the same fixed family tree may contain one species in one world, and two in another—for instance if in one world a certain branch of the tree does not come to occupy a separate ecological niche, and in another world it does. Nor does essential position in a family tree help the essentialist if cladism is true, as LaPorte points out: "according to cladism, a species goes extinct whenever it sends forth a new side species. This is so even if the lineage undergoes no change after sending out the side branch, so that earlier members are indistinguishable from later ones" (p. 103). That is, according to cladism, a thing that is a tiger in this world would not be a tiger in another world, if in that world one of its ancestors had additional offspring that formed a

separate, unrelated lineage constituting a different species. This is so even if the organism's entire ancestry, for the past 3.5 billion years of life on Earth, is held constant. Devitt's final appeal comes in n. 15 (p. 162), where he suggests the possibility of replacing rigid application in his theory with weakly rigid application: "if a term applies to an object in a possible world, then it applies to that object in every possible world in which the object exists and has [the relevant] relational property." Perhaps I don't understand the suggestion, for it seems to me to cast a far wider net than intended. 'Uncle', for example, seems to come out as a rigid applier, because for any uncle, and any world in which he has the relational property of having a sibling who has an offspring, 'uncle' applies to him. One could indeed easily cook up relational properties that would make any general term a weakly rigid applier.

I conclude that LaPorte's objections stand. According to the main classes of contemporary theories concerning the individuation of biological species, no species term is a rigid applier.

It's worth mentioning as well that there are *non*-paradigm cases of (putative) rigid terms that also cause trouble for Rigid Application. Recall Kripke's discussion of Wittgenstein's meter stick (S) early in the first lecture of *Naming and Necessity*. There, Kripke argues that it is the rigidity of 'meter' that explains the contingency of 'S is one meter long':

> [T]here is an intuitive difference between the phrase 'one meter' and the phrase 'the length of S at t_0'. The first phrase is meant to *designate rigidly* a certain length in all possible worlds, which in the actual world happens to be the length of the stick S at t_0. On the other hand 'the length of S at t_0' does not designate anything rigidly… [U]nder certain circumstances, S would not have been one meter long. The reason is that one designator ('one meter') is rigid and the other designator ('the length of S at t_0') is not. [pp. 55-56; emphasis added]

The reasoning seems to be that S is only contingently one meter, because 'one meter' rigidly picks out a certain length, which S could have either been longer or shorter than. Yet 'one meter' is not a rigid applier, precisely because things (like S) that are one meter could have been longer or shorter. Furthermore, I take it that Kripke's explanation of the contingency of 'S is one meter long' is, on the whole, the correct one: there *is* an intuitive difference between 'one meter' and 'the length of S at t_0' and it *is* this difference which explains why 'S, at t_0, is one meter' is contingent whereas 'S, at t_0, is the length of S at t0' is not. If a Rigid Application theorist wants to co-opt this explanation, she's going to need a notion of rigidity in addition to rigid application that captures this intuitive difference—but after we accept an alternate notion of rigidity, it's not clear why we need rigid application anymore, especially when it seems nothing is a rigid applier.

The preceding argumentation has all been to the effect that none of the paradigm cases of rigid general terms turn out to be rigid according to Rigid Application, and thus that the view is INADEQUATE. It's worth reiterating, however, that Rigid Application isn't EXPLANATORY, in that it doesn't explain why such statements as 'water is H_2O' and 'tigers are animals' are necessary, if true. Consider a model with two worlds, w_1 and w_2, and two objects, x_1 and x_2. Suppose that x_1 only exists at w_1 and x_2 only exists at w_2. At w_1, x_1 satisfies F and G, and at w_2, x_2 satisfies F but does not satisfy G. Then both F and G are rigid appliers, for nothing satisfies one of these predicates at one world, but doesn't satisfy it at another. Further, $(x)(Fx \leftrightarrow Gx)$ is true at w_1 (the "actual" world), but it is not necessarily true, since it is false at w_2. Thus, Rigid Application does not allow us to move from a true universally quantified biconditional between two rigid appliers, to its

necessitation—which is presumably what we need to do if we're to explain why 'water is H_2O' is necessary, if true[19].

What we've seen so far then is this. Rigid Application is indeed a natural extension of Kripke's notion of singular term rigidity to the general term case. However, none of the paradigm cases of rigid general terms (secondary quality predicates, natural phenomenon terms, chemical kind terms, and terms for species and higher biological taxa) are rigid appliers (potential exception: 'light' as used by scientists). Furthermore, *even if they were*, the fact that they were would not explain the most striking results of Kripke and Putnam's work, namely that theoretical identity statements involving such terms are necessary, if true. I propose then that if there is a viable alternative account of general term rigidity, it should be adopted. And there is, and it should: rigid expression.

3. Rigid Expression

The view I endorse is that a general term is rigid iff it is a rigid expresser, and I repeat the definition of the latter term for the reader's convenience:

> [Rigid Expression] A general term G is a rigid expresser iff if it expresses the property P relative to the actual world, then it expresses P relative to all worlds.

(It's worth reiterating that the world-relativity here is relativity to a world of evaluation, not the world parameter of the context of utterance.)

There are three main lines of objection to Rigid Expression in the literature. First, it is maintained that all and only non-descriptive terms are rigid, whereas many descriptive terms are rigid expressers: thus, rigidity is not rigid expression. Second, it is

[19] Soames (2002) makes the same point on pp. 257-259. Devitt (2005) concedes the point on p. 152.

maintained that all and only natural kind terms are rigid, whereas many non-natural kind terms are rigid expressers: thus, rigidity is not rigid expressions. Finally, some have claimed that "rigid expresser" is a trivial notion—not merely that it can't provide a descriptive/ non-descriptive term distinction or a natural kind/ non-natural kind term distinction, but that it can't do any work at all in our semantic theory.

Schwartz (2002) suggests the first line of attack: "Clearly there is an important difference between natural kind terms like 'gold' and nominal kind terms like 'bachelor'– and isn't this difference based on the rigidity of the one and nonrigidity of the other?" (p. 266). This seems to be a problem for Rigid Expression, for it would seem to count both 'gold' and 'bachelor' as rigid—for instance, 'bachelor' expresses the same property relative to every world. We can see this because 'John is a bachelor' is true, at an arbitrary world w, iff John has *the property of being a bachelor* in w. I take Soames to be making a similar point when he says that extending rigidity to all terms is "problematic"[20] because "Kripke wanted to distinguish natural kind predicates like *is gold* and *is a tiger* from ordinary descriptive predicates such as [*is a philosopher, is a bachelor*, etc.]" (2002, p. 260).

The Soames/ Schwartz suggestion, that rigid general terms should be all and only the natural kind terms, seems off-base. After all, whether a general term is a natural kind term or not depends on what property it expresses. If it expresses a natural kind property, it is a natural kind term; otherwise, it isn't. But whether a singular term is a rigid designator does not depend on what object it designates. Any object can be designated

[20] In fairness to Soames, he says that *interpreting Kripke* as holding Rigid Expression is problematic for this reason, and this is consistent with Soames not seeing any problem at all with 'gold' and 'bachelor' both being rigid.

rigidly, and any object can be designated non-rigidly[21]. The natural kind term/ non-natural kind term distinction is a distinction between terms with one type of semantic value and terms with a different type of semantic value. But the rigid term/ non-rigid term distinction is a distinction between terms that bear one type of relation to their semantic value (a persistent one) and those that bear a different type of relation to their semantic value (a non-persistent one). It would be rather surprising if the two distinctions coincided. And this very point shows, contra Schwartz, that the rigid/ non-rigid distinction is not needed to capture the difference between natural kind terms and other general terms. Natural kind terms are those that express properties which are individuated by their microstructure. Other general terms don't.

There is another reason why it would be surprising if all and only natural kind terms were rigid general terms. The rigidity of 'tiger' and 'animal,' we are assuming, is supposed to explain why 'all tigers are animals' is necessary if true. One might have thought then that the necessity of 'all hammers are tools' is also to be explained by the rigidity of 'hammer' and 'tool.' Or, at least, a theory that could produce such an explanation would be better than a theory that couldn't. So there's no sense in prejudging the case against theories that would try such explanations, as Soames and Schwartz seem to want to do.

Devitt (2005) joins me in thinking that Soames and Schwartz are making "a mistake." He says:

> [I]t is a mistake to think that the primary task of the rigidity distinction is to distinguish natural kind terms from nominal kind terms. The primary task is to

[21] "[R]igidity is a semantic claim about a designator, not a metaphysical claim about the essence of what is designated" (Sullivan 2007, p. 6b). QFT.

> distinguish kind terms that are not covered by a description theory from ones that are. [p. 154]

This task is equally problematic for Rigid Expression, since it might well be thought that 'bachelor' is descriptive, and thus should be non-rigid (if the primary task of rigidity is distinguishing what's descriptive from what isn't), but it comes out rigid on Rigid Expression.

On the one hand, I have no compunction in agreeing with Devitt. Since I've argued in Chapter 1 that all general terms are non-descriptive, and since I'm arguing now that they're all rigid, Rigid Expression does the work expected of it. A general term is rigid, according to me, iff it is non-descriptive. But on the other hand, it is clearly Devitt who is mistaken. Jeff King (p.c.) has pointed out to me that in a technical sense, definite descriptions are indeed "not covered by the description theory." In particular, they don't satisfy Kripke's six descriptivist theses (1980, pp. 54-55), because what they denote is not determined by properties that speakers associate with them, but rather is determined by the meanings of their parts and the compositional rules of the language. So if Devitt is right, the purpose of rigidity for singular terms is to class names and definite descriptions together as rigid, and set them apart from descriptive names (like 'Julius,' maybe?). But that just isn't right: 'the inventor of bifocals' is non-rigid, if anything is.

A more charitable reading of Devitt's suggestion is that he means definite descriptions to count automatically as "covered by the description theory" (they are after all called definite *descriptions*), and therefore count automatically as non-rigid. However, rigidity doesn't distinguish between names and definite descriptions, as pointed out by, say, Kripke (1980). Actualized definite descriptions (like 'the actual teacher of

Alexander') and certain descriptions designating necessarily existing objects (like 'the successor of 0') are rigid designators, just like names.

The upshot is this. There are rigid definite descriptions and non-rigid definite descriptions. If you count definite descriptions as not being covered by the description theory, since they don't satisfy Kripke's descriptivist theses, then Devitt's suggestion nets you the wrong result. If you count definite descriptions as being covered by the description theory, because they're after all *descriptions*, Devitt's suggestion nets you the wrong result. I suppose Devitt could say 'the teacher of Alexander' is descriptive, whereas 'the actual teacher of Alexander' is non-descriptive, but I don't know on what grounds he'd propose to do so. Thus I conclude that if general term rigidity is indeed to be an extension of Kripke's notion of singular term rigidity, then contra Devitt it should not have as its purpose distinguishing descriptive general terms (if there are any) from non-descriptive ones.

The most serious charge against Rigid Expression is that according to it, every general term is trivially rigid, because every term trivially expresses the same property relative to every possible world. This is certainly true for some understandings of 'property.' For example, Carnap took properties to be intensions (1988, pp. 18-19). The intension of a predicate P is a function that maps a possible world w to the extension of P relative to w (construed as a world of evaluation). By definition, then, no predicate can have different intensions relative to different possible worlds. For suppose P has intension I_1 relative to w_1 and intension I_2 relative to w_2, where $I_1 \neq I_2$. If I_1 and I_2 are distinct, they must differ in the extension they assign to some argument w, i.e. $I_1(w) \neq I_2(w)$. But predicates have but one extension relative to each world, so either $I_1(w)$ is not

the extension of P at w, or $I_2(w)$ isn't. If $I_1(w)$ is not the extension of P in w, then by definition I_1 is not the intension of P; similarly if $I_2(w)$ is not the extension of P in w, then by definition I_2 is not the intension of P. Thus predicates have at most one intension. It's of little use to talk of predicates having intensions *relative to worlds*, but if we allow such talk, we must admit that trivially, i.e. purely as a matter of what it is to be an intension, each predicate has the exact same intension relative to each world. If we call the semantic relation a predicate bears to its intension the intending relation, and call the semantic persistence of the intending relation rigid intending, then it is a trivial truth that all predicates are rigid intenders[22].

But there is no need to accept the identification of properties with intensions. Even a relatively non-committal account of properties—properties are ways that things can be—can avoid the triviality charge, or so it seems. There is nothing in the definition of "way a thing can be" that requires that general terms express the same way for things to be relative to each possible world, in the way that the mere definition of an intension requires terms to have the same intensions relative to each world.

Several philosophers, however, hold that even on a non-intensional reading of 'property,' Rigid Expression is trivial. I wish to dispute the charge. First, however, let us examine it. Soames (2002) says:

> Nor will it do to say that a predicate is rigid iff there is a unique property which it stands for that determines its extension at each possible world. There is, it could be argued, such a property in the case of natural kind predicates like *cow* and *animal*—namely, the property of being a cow and the property of being an animal. However, the same could be said for any predicate; for any predicate F,

[22] What about 'is an actual philosopher'—does this have distinct intensions relative to different worlds? No. The character of an expression maps a context of utterance (which contains a world parameter) to an intension. Relative to different contexts of utterance, an expression with a non-constant character may indeed have different intensions. But if we fix the context of utterance (as I intend to do in my argument), no word (by definition) may have different intensions relative to different *worlds of evaluation*.

> and any world w, the extension of F with respect to w is the set of things that have, in w, the property expressed by ***being an F***. But there is no point in defining a notion of rigidity for predicates according to which all predicates turn out, trivially, to be rigid. [pp. 250-251]

And Stephen Schwartz (2002) says:

> [T]his solution [Rigid Expression] is unsatisfactory because, among other things, it extends the privilege of rigidity to just about all general terms (Schwartz, 1980). 'Bachelor' will designate the same kind – the same marital status – in every possible world in which it designates. Likewise for other nominal kind terms. They all turn out to be rigid. To some this result would be welcome, but it seems to me to lose all the ground gained. Rigidity has lost its exclusivity, like a club of which all are automatically members, and thereby its interest… The basic problem is that this proposed solution trivializes rigidity. [p. 266]

The thought here seems to be that if every general term is rigid, then rigidity (for general terms) cannot play any role in explaining semantic phenomena, and thus there is "no point" in working with such a concept of rigidity.

The first thing I want to do is concede one aspect of the charge, before denying the other. According to me, all general terms are rigid. Each general term expresses a property, and it expresses the same property relative to every possible world. 'Cow' rigidly expresses the property of being a cow, 'philosopher' rigidly expresses the property of being a philosopher, and so on. Rigidity is not an exclusive club. What I want to deny, however, is that this somehow trivializes the notion of rigidity. All general terms are rigid, yes, but this does not follow from the definition of rigid expression alone. Each general term expresses the same property relative to each possible world, but this might not have been so. Rigidity is not an exclusive club, but it could have been

Others who endorse rigid expression are inclined toward a different position. Linsky (1984), LaPorte (2000), and Salmon (2005b) are at great pains to show that even

if Rigid Expression is correct, not all predicates are rigid. LaPorte (2000) clearly sees this as a necessary task in defending against the triviality objection:

> Of course, that the account at issue [Rigid Expression] can show how rigidity might be applicable to kind designators does not assure its acceptability. If it is to be satisfactory, the account must allow for terms relevantly like 'water' and 'H_2O' to come out rigid. And, presumably, it *must* make other expressions designating kinds come out non-rigid. [pp. 294-295, emphasis added]

The paradigm case of a non-rigid predicate is supposed to be something like 'the color of the sky,' as in the sentence "my true love's eyes are the color of the sky." Linsky (1984) and Salmon (2005b, p. 124) hold that the expression is a second-order definite description, roughly equivalent to "the unique property F such that F is a color property and the sky is F." The designatum of this description evaluated at w_1, where the sky is green, is the property of being green; at w_2, where the sky is orange, is the property of being orange. Hence, it is not a rigid designator: it designates different properties relative to different worlds.

There are objections to this treatment in the literature, but I think Rigid Expression theorists have handled them well enough. The most current runs along these lines: 'the color of the sky' is not non-rigid after all. Instead, it rigidly designates one and the same property relative to each possible world, namely, the property of being the color of the sky. This is the property that green things have, in worlds where the sky is green, and the selfsame property that orange things have, in worlds where the sky is orange (cf. Schwartz, 2002, Section III). And the clever reply is this: if 'the color of the sky' is indeed a rigid designator, it should be equivalent to its actualization, 'the actual color of the sky.' But 'grass is the color of the sky' is true when evaluated at a world where the

sky is orange and grass is orange, whereas ‘grass is the actual color of the sky’ is false at such a world (Linsky, 2006; Martí & Martínez-Fernández, 2010, p. 50).

Nevertheless, there’s a much more serious objection to the Linsky-Salmon treatment that is not in the literature. On a neo-Russellian account of definite descriptions, descriptions have as their semantic value quantifiers, not what they denote (that is, whatever uniquely satisfies the semantic value of the NP complement to ‘the’). On a uniform treatment of rigidity, we should say that an expression is rigid iff it has the same semantic value relative to every possible world. So names are rigid, because each refers to the same object relative to every possible world; but definite descriptions (including second-order descriptions like ‘the color of the sky’) are also rigid, because their semantic value is the same quantifier relative to every possible world. The Linsky-Salmon treatment only delivers non-rigid designators because it assumes a non-uniform account of designation: names designate their semantic value, but definite descriptions designate whatever uniquely satisfies the semantic value of the NP complement to ‘the.’

Of course, this assumption is bona fide Kripke, but it was never clear that it was justified in Kripke either. It’s one thing to claim that the falsehood of ‘Aristotle might not have been the D’ for most D’s (i.e. not including ‘person identical to Aristotle,’ ‘son of Aristotle’s mother,’ etc.) shows that ‘Aristotle’ is not synonymous with a description; it’s quite another to claim that the truth of ‘the teacher of Alexander the Great is such that he might not have been the teacher of Alexander the Great’ shows that ‘the teacher of Alexander the Great’ is non-rigid. The falsehood of the latter sentence does not show that ‘the teacher of Alexander’ may not be substituted with co-designators salva veriatate in all modal contexts, *unless it is assumed that what it designates is an individual and not a*

quantifier. For substitution of an equivalent quantifier, e.g. ‘the teacher of Alexander III of Macedon,’ does not result in a change in truth-value for any sentence in which ‘the teacher of Alexander the Great’ occurs.

Let’s try to get clearer on what the triviality objection amounts to. One way of interpreting the Soames quote above is that Soames is claiming that rigid expression is a trivial property of general terms because, as a matter of fact, all general terms are rigid expressers. That is, the trivial properties of things in domain D are all those properties had by every member of D. This seems like a bad interpretation, because it is no mark against a property that it be trivial in this sense. According to many, every complex expression is compositional (has its meanings determined by its parts and their mode of combination). And yet, there is a point in defining compositionality, even if all complex expressions are compositional.

A second way of interpreting ‘trivial’ here is taking the trivial properties of things in a domain D to be all those properties that every member of D can be known to possess a priori. Thus, the objection would be that rigid expression is a trivial property of general terms, because it’s a priori that every general term is a rigid expresser.

This objection fails, though, because according to some theorists, there are indeed non-rigid general terms. There is a way of reading “Mad Pain and Martian Pain” (Lewis, 1980) in which what Lewis is claiming is that ‘pain’ picks out one property relative to some indices, and picks out a different property relative to other indices. This reading is tricky, for although Lewis explicitly states that according to him, ‘pain’ is non-rigid, he also seems to be assuming Rigid Application, for he says: “the concept and name of pain contingently apply to some neural state at this world, but do not apply to it at another” (p.

218b) and takes this to decide the non-rigidity of 'pain.' Nevertheless, he does say that 'pain' takes on different "senses" at different indices, as when he says "The madman is in pain in one sense, or relative to one population. The Martian is in pain in another sense, or relative to another population" (p. 221a)[23] On this reading, Lewis denied that all general terms were rigid expressers, and therefore it is not a priori that Rigid Expression is true unless it's a priori that Lewis was wrong.

Even if this interpretation of Lewis is implausible, and even if no other theorist does or would hold such a view, still I think the objection that the properties that can be known a priori to be possessed by all members of domain D are trivial in a bad sense is misguided. This is simply on account of the fact that a priori knowledge is often hard-won, and sometimes far more difficult to obtain than run of the mill knowledge. Any logical system in which one can derive the Peano axioms is incomplete, but this fact is hardly trivial. It strikes me that the same is true of the claim that all general terms are rigid.

Perhaps these two interpretations of 'trivial' are incorrect or uncharitable. Nevertheless, there's strong reason to suppose that no sense of triviality that makes Rigid Expression a trivial thesis is a triviality worth avoiding. Recall that according to Kripke, all names are rigid, and that according to Kaplan, all simple demonstratives and indexicals are rigid. This entails that all singular terms are rigid, on the further assumption that neither definite descriptions (Neale, 1990) nor complex demonstratives (King, 2001) are singular terms. Call this suite of views N = ST (for "the names are all

[23] A note on populations: in context, it is quite clear Lewis is using them here in the same way as, and often as proxy for, possible world indices of evaluation. He introduces these cases with the assertion: "If a nonrigid concept or name applies to different states in different possible cases, it should be no surprise if it also applies to different states in different actual cases" (p. 219a).

and only the singular terms"). N = ST seems to be in the same boat as Rigid Expression. Both views have a notion of rigidity (rigid reference, rigid expression) that applies equally and indiscriminately to every member of their respective domains (singular terms, general terms). So unless it can be maintained that N = ST trivializes the standard account of rigidity for singular terms, it's hard to see how it can be maintained that Rigid Expression trivializes rigidity for general terms.

What I think is the heart of the triviality objection is the worry that extending rigidity to all general terms will neuter its explanatory value. We've already seen that rigid expression cannot help us in distinguishing descriptive general terms from non-descriptive general terms, or in distinguishing natural kind terms from non-natural kind terms. But the important question is whether Rigid Expression is EXPLANATORY, that is, whether it can explain why such claims as *water is H_2O* and *tigers are animals* are necessary. If it's EXPLANATORY, then it's explanatory, and so not trivial. In the next section, I argue that Rigid Expression is indeed EXPLANATORY.

4. Essentialist Conclusions

An account of general term rigidity is EXPLANATORY iff it can explain the necessity of 'water is H_2O' and 'tigers are animals' in roughly the same way that singular term rigidity explains the necessity of 'Hesperus is Phosphorus.' It is appropriate then to investigate just what role rigidity is supposed to be playing in explaining the necessity of 'Hesperus is Phosphorus.'

Consider the following argument:

1. ‘Hesperus’ and ‘Phosphorus’ are rigid designators
2. Hesperus = Phosphorus.
3. (x)(y)(x = y → □ x = y)
4. □(Hesperus = Phosphorus)

The conclusion (4) will not be true if any of the premises (1)-(3) are false. For instance, consider an analogous argument, where ‘Hesperus’ and ‘Phosphorus are replaced with non-rigid designators:

2’. The first postmaster general = the inventor of bifocals
3’. (x)(y)(x = y → □ x = y)
4’. □(the first postmaster general = the inventor of bifocals)

The invalidity of (2’)-(4’) (on the narrow-scope construal of the descriptions) seems to clearly show that it is the rigidity of the singular terms in the argument that licenses the substitution of identicals from (2) to (3)[24]. After all, if relative to some worlds ‘Hesperus’ designated Mars, then though it should be true that in the actual world, Hesperus = Phosphorus, it might fail to be true that necessarily, Hesperus = Phosphorus, in particular, in those worlds relative to which ‘Hesperus’ designated Mars while ‘Phosphorus’ designated Venus.

But premise (3) is also integral to the argument. Suppose it were false, and in some worlds, Hesperus was diverse from itself. Then we could argue:

1’’. ‘Hesperus’ and ‘Phosphorus’ are rigid designators
2’’. Hesperus = Phosphorus
3’’. ◊ (Hesperus ≠ Hesperus)
4’’. ◊ (Hesperus ≠ Phosphorus)

That is, the truth of ‘necessarily, Hesperus = Phosphorus’ is dependent *both* on the rigidity of both ‘Hesperus’ and ‘Phosphorus’ *and* on the essentialist assumption that everything is necessarily self-identical.

[24] And since bound variables are rigid designators, we can see why (1’)-(4’) is valid on the wide-scope construal of the descriptions.

The role of essentialist assumptions in such reasoning can be made clearer by considering a different example. Suppose, as in (7), that mereological fusions have their parts essentially.

5. ‘Abbott’ and ‘Abbott + Costello’ are rigid designators
6. Abbott ≤ Abbott + Costello
7. (x)(y)(x ≤ y → □ x ≤ y)
8. □ (Abbott ≤ Abbott + Costello)

Here, although (7) does much of the heavy lifting, (5) is still required for the argument to go through. For suppose ‘Sam’ is introduced as a non-rigid name that relative to each world, designates the fusion of the two members of the greatest comedy duo in that world. Then (6’)-(8’) is invalid:

6’. Abbott ≤ Sam
7’. (x)(y)(x ≤ y → □ x ≤ y)
8’. □ (Abbott ≤ Sam)

In particular, Abbott might be part of the greatest comedy duo in the actual world, satisfying (6’), but not part of the greatest comedy duo in some other world, falsifying (8’).

This then is my read on rigidity’s relevance for the necessity of ‘Hesperus = Phosphorus’ and ‘Abbott ≤ Abbott + Costello.’ First, there are certain metaphysical laws concerning individuals such as (x)(y)(x = y → □ x = y) and (x)(y)(x ≤ y → □ x ≤ y). Second, any sentence that expresses an instance of these laws will be true—that’s what it means for them to be laws. But *only* sentences where rigid designators are substituted for the bound individual variables in the law express instances of that law. No fact about language makes it the case that objects are necessarily self-identical. Yet had our language been different, and had our singular terms been non-rigid, statements like a = b & ~□ a = b would come out true, since they would mean (in our idiom) “a = b and

there's some world w where c ≠ d". The rigidity of singular terms in our idiom makes such statements come out false, and makes it legitimate to substitute names for the bound variables in Leibniz's Law.

Furthermore, I claim that rigid expression, the property a general term has when it expresses the same property relative to every world, plays exactly the same role as rigid designation in explaining certain necessary statements concerning properties. Consider the following argument:

9. 'Groundhog' and 'woodchuck' are rigid expressers
10. Groundhog ≡ woodchuck
11. (X)(Y)(X ≡ Y → □(z)(Xz ↔ Yz))
12. □(z)(groundhog(z) ↔ woodchuck(z))

Here, I use the symbol '≡' to denote property identity, so that (10) is to read "the property of being a groundhog is numerically identical to the property of being a woodchuck' and (11)[25] is to read "any two numerically identical properties are necessarily coextensive." Again, although (11) does the heavy lifting, (9) is required for the argument to be valid. For suppose 'schwoodchuck' is a non-rigid expresser, that expresses the property of being a woodchuck relative to the actual world and the property of being a ham sandwich relative to every other world. (10')-(12') are then clearly invalid:

10'. Woodchuck ≡ schwoodchuck
11'. (X)(Y)(X ≡ Y → □(z)(Xz ↔ Yz))
12'. □(z)(woodchuck(z) ↔ schwoodchuck(z))

In particular, every world besides the actual world is a counterinstance to (12'), but (10') and (11') are both true.

[25] (11) looks distressingly similar to (one direction of) Frege's Basic Law V. However, the similarity is unproblematic, as I'm assuming a background logic that doesn't allow the formation of expressions like 'being a property not had by yourself' (i.e. λX. ∃F X ≡ F & ~FX. Here ~FX is ill-formed. Essentially, the typing does all the work.)

I must note that I am *not* claiming that the English sentence ‘groundhogs are woodchucks’ expresses the proposition that groundhog ≡ woodchuck. Some who have endorsed Rigid Expression do hold this (Martí & Martínez-Fernández, 2010, pp. 56-57), though I take it to be unnecessary and implausible. It’s implausible because it involves positing a new sense of English ‘be,’ as “fire trucks are red’ clearly doesn’t express the proposition that fire truck ≡ red. Furthermore, this ad hoc sense of ‘be’ is insufficient for the task at hand, since it won’t explain why “tigers are animals” is necessarily true (as it doesn’t mean tiger ≡ animal). Finally, the proposal is unnecessary, as the mere fact that groundhog ≡ woodchuck guarantees the necessity of “groundhogs are woodchucks” *even if* the latter sentence only means (z)(groundhog(z) → woodchuck(z)) and nothing stronger.

It may well be objected that (10) groundhog ≡ woodchuck is not a truth, not a truth we have access to, nontrivially essentialist in character, or really just the conclusion (12) □(z)(groundhog(z) ↔ woodchuck(z)) in disguise. I certainly think (10) is true. If properties were concepts, and as such satisfied Frege’s individuation criterion (identical properties are a priori knowable to be so; distinct properties are a priori knowable to be so), then (10) would be false, since not knowable a priori. But properties are not concepts (as I’m using the term ‘property’), and two general terms can express the same property without our knowing, and in some cases perhaps without our being able to know, that this is so.

But how can it be known that the property of being a groundhog and the property of being a woodchuck are numerically one and the same? I suppose in much the same way that it can be known that Samuel Clemens and Mark Twain are one and the same.

They look and quack like each other. Same handwriting. Same fingerprints. Same DNA. Likewise, groundhogs and woodchucks are suspiciously similar. They're both marmots; both like alfalfa; both hibernate; both get scared by their shadows, especially when spring will be late. Our data may underdetermine our theory, but that the property of being a groundhog is identical to the property of being a woodchuck is as good a guess as any, and one I'd put money on.

Finally, premise (10) is not an underhanded way of sneaking our conclusion into our premise set. Our conclusion (12) simply asserts that 'groundhog' and 'woodchuck' have equivalent intensions. But this is not sufficient for the properties they express to be identical: 'triangular' and 'trilateral' have equivalent intensions, but express distinct properties. Furthermore, (10) is a nonmodal, empirically discoverable fact, on a par with Hesperus = Phosphorus. (12) is something more: it tells us how worlds other than the actual world must be like.

So (9) is true, given our theory of reference. (10) is a true, nonmodal, empirically discoverable fact. And (11) is just plain solid metaphysics: identical properties have identical intensions. What's more, (9)-(11) entail (12). No better candidate explanation for the necessity of 'groundhogs are woodchucks' has ever been proposed; this looks like it.

(It might be thought suspicious that when I issued my challenge to myself, to explain why identities between rigid general terms were necessary, if true, I used the example '$(x)(water(x) \leftrightarrow H_2O(x))$' and yet when I proceeded to an explanation, my example changed to '$(z)(groundhog(z) \leftrightarrow woodchuck(z))$.' And this was indeed a bit of sleight-of-hand on my part, for which I apologize. Let me explain. The thesis of this

paper is that all general terms are rigid, and by 'term' I mean to exclude complex expressions. Up to now, I'd been happy pretending 'H_2O' was a term, but by my own lights, it patently isn't. I do have some remarks about complex expressions though; you can find them in Section 6.)

The work of the Rigid Expression theorist is not done here. For we must also capture the fact that it is the rigidity of 'tiger' and 'animal' that accounts for why 'tigers are animals' is necessary, and this can't be done with Leibniz's Law for general terms, since patently the property of being a tiger is *not* the property of being an animal. Martí & Martínez-Fernández, who hold a view largely similar to my own, handle the case as follows: they claim to adopt the EXPLANATORY desideratum from Soames, while nevertheless quietly dropping the half of the desideratum requiring an explanation for the necessity of statements like 'tigers are animals' (see Soames (2002) p. 263 and cf. Martí & Martínez-Fernández p. 47). But I'd rather not avoid the difficult case by stipulating that I don't have to consider it.

In fact, an argument along the same lines of (9)-(12) can be given to show that 'tigers are animals' also expresses a necessary truth. Let '$\leq$' denote the relation property A bears to property B when objects that are A are B in virtue of being A. For instance, a particular shade of red is red in virtue of being the particular shade it is; a tiger is an animal in virtue of its being a tiger; a hammer is a tool in virtue of its being a hammer. On the other hand, a shade of red is not red in virtue of being on the wall; a tiger is not an animal in virtue of being striped; and a hammer is not a tool in virtue of being possessed by an office clerk. So magenta $\leq$ red; tiger $\leq$ animal; and hammer $\leq$ tool; but it is false,

for instance, that striped $\leq$ animal. It seems intuitive that when $A \leq B$, $\Box(z)(Az \rightarrow Bz)$. So now consider[26]:

13. 'Tiger' and 'animal' are rigid expressers
14. Tiger $\leq$ animal
15. $(X)(Y)(X \leq Y \rightarrow \Box(z)(Xz \rightarrow Yz))$
16. $\Box(z)(\text{tiger}(z) \rightarrow \text{animal}(z))$

Hopefully it's clear by now that (13) is a needed premise in the argument.

I don't think there's anything particularly fishy about the relation $\leq$. Properties certainly stand in logical relations to one another. Assume that the property of being large is identical, upon metaphysical analysis, to the property of being either big or tall. Then someone who has the property of being tall will, as a point of logical necessity, have the property of being large. Or suppose the property of being colored is identical, upon metaphysical analysis, to the property of reflecting photons with wavelengths between 390 and 750nm, and that the property of being red is identical, upon metaphysical analysis, to the property of reflecting photons with wavelengths between 635 and 700nm. Then some surface which has the property of being red will, as a point of logical necessity, have the property of being colored. I think the biological cases will be similar: to be an animal is to have an evolutionary history of type H; to be a tiger is to have a history of type H*; and having H* logically entails having H. How exactly properties (and not linguistic representations of them) stand in logical relations is perhaps somewhat mysterious, but I don't think it is to be doubted, and that is all that is needed here.

[26] Jeff King (p.c.) worries that (15) is false because of cases like this: I live in the U.S. in virtue of living in New Jersey. Yet it's not necessary that New Jersey residents live in the U.S., for Jersey could secede from the union. I'm inclined to think the first premise in the argument is wrong: I live in the U.S. not in virtue of living in New Jersey full stop, but in virtue of living in New Jersey *and New Jersey's bearing the appropriate relation to the federal government*. And there's no world in which that's true but I don't live in the U.S.

The reader might object that □(z)(groundhog(z) ↔ woodchuck(z)) was supposed to follow from the weaker assumption (z)(groundhog(z) ↔ woodchuck(z)) perhaps together with the fact that 'groundhog' and 'woodchuck' are kind terms of the same type, not the stronger groundhog ≡ woodchuck. Second, it can be similarly objected that □(z)(tiger(z) → animal(z)) was supposed to follow from the weaker assumption (z)(tiger(z) → animal(z)), again with assumptions about the kind term status of 'tiger' and 'animal', rather than from tiger $\leqq$ animal.

Consider then the objection that (z)(groundhog(z) ↔ woodchuck(z)), together with the assumption that 'groundhog' and 'woodchuck' are rigid predicates, should entail, with no other hypotheses, □(z)(groundhog(z) ↔ woodchuck(z)). The motivation I'm imagining someone has who would press this objection is something like this. "Look, Kripke said that theoretical identity sentences involving rigid designators are necessary, if true. He gives the example 'water is H_2O.' For a multitude of reasons, the logical form of 'water is H2O' is (z)(water(z) ↔ H_2O(z)). So if that's true, and if 'water' and 'H_2O' are indeed rigid designators, then it had better follow *from these facts alone* that □(z)(water(z) ↔ H_2O(z)). If you introduce auxiliary or stronger hypotheses to show that water is necessarily H_2O, then you're not explaining what Kripke was talking about." (Again, I'll keep to examples involving terms, not complex expressions—which latter are dealt with in Section 6. I recognize that Kripke's "theoretical identity statements" were rarely identities between terms, but in the interests of not conflating distinct issues, my way will have to be *the* way, for the moment.)

There are two things I should like to say in reply to my hypothetical interlocutor. The first is, that she can't have what she wants. We've been through the only three

possible accounts of rigidity. Rigid extension and rigid application don't even *try* to explain the necessity of 'groundhogs are woodchucks.' If there's any hope of explaining this, the arguments I've given are it. Second, it would be terrible if she could have what she wants. Suppose for instance that all and only ravens are black. Kripke plausibly argues that both 'raven' and 'black' are rigid designators. So if $(x)(Fx \leftrightarrow Gx)$ and the rigidity of 'F' and 'G' are sufficient for demonstrating that $\Box(x)(Fx \leftrightarrow Gx)$, then it should follow, on such a supposition, that necessarily ravens are black. Frankly, that's absurd. Some additional hypothesis is needed. I have characterized the hypothesis as: the property of being a groundhog and the property of being a woodchuck are identical, and identical properties have necessarily identical extensions. These hypotheses are true, I've argued, and pretty much unobjectionable.

Finally, I should point out that the explanations considered above extend to nonnatural kind terms as well, so that we can explain why 'all and only water closets are bathrooms' and 'all hammers are tools' are necessary, if true. The explanations look like this:

17. 'WC' and 'bathroom' are rigid expressers
18. $WC \equiv$ bathroom
19. $\underline{(X)(Y)(X \equiv Y \rightarrow \Box(z)(Xz \leftrightarrow Yz))}$
20. $\Box(z)(WC(z) \leftrightarrow bathroom(z))$

21. 'Hammer' and 'tool' are rigid expressers
22. Hammer $\underline{\leq}$ tool
23. $\underline{(X)(Y)(X \leq Y \rightarrow \Box(z)(Xz \rightarrow Yz))}$
24. $\Box(z)(hammer(z) \rightarrow tool(z))$

This generality is to be appreciated.

5. Yes, But Why?

If I'm right, then general term rigidity is rigid expression, and since all general terms are rigid expressers, all general terms are rigid. One might rightly ask why this should be so. Indeed, even setting aside complex predicates, it still isn't obvious why all simple general terms should be rigid expressers. Why *aren't* there lexically simple general terms that say, express the property of being an F relative to the actual world, and the property of being a G (for G ≠ F) relative to any other world? This question deserves to be answered.

A first step toward answering this question is simply realizing that not everything is relative to everything. For example, one and the same individual may fail to be tall, in one world and at one time, relative to the standard of height for NBA centers, but succeed in being tall relative to the standard of height for Danny DeVito impersonators, in that world and time. And yet, it is never the case that one person, in one world and at one time, is pregnant, relative to the standard height for NBA centers, but that very person is also not pregnant, in that world and time, relative to the standard of height for Danny DeVito impersonators. It might well make perfect sense to talk of some person being pregnant, in a world and at a time, relative to some standard of height. But the extra parameter would be a dispensable add-on: we could just as easily talk of being pregnant in a world and at a time *simpliciter*, and omit any mention of standards of height. The insight here is that being pregnant (in a world, at a time) is simply *metaphysically independent of* standards of height.

Perhaps then, expression (the relation a predicate bears to the property that determines its extension) is a relation that is metaphysically independent of possible worlds. If this were so, then the fact that some general term G expressed the property P

relative to every world would just amount to the fact that G expressed the property P—talk of expressing *relative to worlds* could be trivially replaced by talk of expressing simpliciter, just as talk of being pregnant *relative to standards of height* is trivially replaced by talk of being pregnant simpliciter. That is, it would follow directly from one's being an expresser that one was a rigid expresser.

The analogue of this view for certain rigid singular terms seems to have been hit upon twice independently[27]. David Kaplan: "the referent, in a circumstance, of a directly referential term is simply *independent* of the circumstance and is no more a function (constant or otherwise) of circumstance, than my action is a function of your desires when I decide to do it whether you like it or not" ("Demonstratives," p. 756). And Gareth Evans: "a referring expression does not designate the same thing with respect to each possible situation; it simply designates, and the truth value of any sentence containing it depends upon what, if anything, it designates" ("Reference and Contingency", p. 192). And indeed, something like this picture is suggested by Kripke himself. In the final footnote of the introduction to the 1980 printing of *Naming and Necessity*, he draws a distinction between de jure and de facto rigidity:

> I ignore [in the monograph] the distinction between '*de jure*' rigidity, where the reference of a designator is *stipulated* to be a single object, whether we are speaking of the actual world or a counterfactual situation, and mere '*de facto*' rigidity, where a description 'the x such that Fx' happens to use a predicate 'F' that in each possible world is true of one and the same unique object… clearly my thesis about names is that they are rigid *de jure*. [p. 21 n. 21][28]

[27] "Demonstratives" was written in 1977, before Evans' paper. Evans' first citation to "Demonstratives" doesn't appear until 1981 (in "Understanding Demonstratives"), three years after "Reference and Contingency" was published, so it is entirely likely that in writing the former work, he had not seen Kaplan's paper (which was unpublished at the time).

[28] Scholarly nitpick: Martí (2003) p. 134 quotes this passage and claims that the emphasis on 'stipulated' is added by her; she is wrong on this point—the emphasis is in the original.

Some commentators, such as McGinn (1982), take this to mean that de jure rigid designators are not to be seen as expressing constant functions from worlds to semantic values (this is what de facto rigid designators do), but rather as simply delivering a semantic value (an object, not a function from worlds to objects).

If a general term merely contributes a property to propositions expressed by sentences in which it occurs, rather than something that *determines* a property (like a description, or a function from worlds to properties), then it will be a rigid expresser. We might wonder, though, *why* general terms merely contribute properties, rather than things that determine properties, to propositions. And here an intriguing line of thought presents itself: since there are a myriad of reasons why natural language expressions should not be descriptive (see previous chapter), perhaps those reasons themselves explain why there aren't, in addition to rigid general terms, general terms that are non-rigid. The line of thought I'm gesturing at is something like this: There's ample explanation for why general terms are non-descriptive; non-descriptive terms are all rigid; therefore, there's ample explanation for why general terms are rigid.

I think this line of thought is indeed intriguing, but I'm inclined to doubt the second premise, that non-descriptive terms are all rigid. Soames gives the argument for it as follows:

> Since names are nondescriptional, the referent of a name at a world is not semantically determined by the satisfaction of any descriptive condition at that world; thus, there is no semantic mechanism by which the reference of a name might change from world to world. Instead, the referent of the name is initially fixed at the actual world (for example by a historical chain of use or by a reference-fixing description) and, once fixed, there is no provision for it to vary from world to world. [2002, p. 264]

This *can't* be right. Recall that for a term (let's say a name) to be descriptive, its referent must be determined by a descriptive condition *that the speaker associates with the name*. So it's compatible with a name's being non-descriptive that its referent is determined by a descriptive condition, just not one any speaker does or is required to (on pain of incompetence) associate with the name. For instance, an entirely conceivable semantic theory might say that the referent of 'Saul Kripke' is determined by the description 'the most natural object that satisfies most of the community's "Saul Kripke"-beliefs' even though no competent speaker is required to believe this, or even to know that there are distinctions of naturalness among various existents. In such a case, 'Saul Kripke' is nondescriptive, in that there is no condition that any speaker associates with the expression that determines its referent, but there is nevertheless a provision in the semantics that allows it to vary its referent from world to world.

So we can't explain why all names or general terms, say, are de jure rigid ("directly referring," in Kaplan's sense), merely by giving an explanation for why they're all non-descriptive. Something else must be said to rule out cases where names or general terms are non-descriptive, but still have their contents determined by descriptive conditions (ones that no-one *associates* with the expression).

At bottom, I think the answer is as simple as this. Our reasoning about modal and counterfactual circumstances is governed by various principles, among them Leibniz's Law and all its substitution-instances involving singular terms. If we adopt, say, an informational metasemantics, then 'Aristotle' refers to Aristotle outside of modal contexts because 'Aristotle'-involving sentences carry information about Aristotle's properties, not someone else's. Further, 'Aristotle' refers to Aristotle inside of modal

contexts, because our adherence to Leibniz's Law guarantees that modal 'Aristotle'-involving sentences carry information about Aristotle's modal properties, not someone else's. If, on the other hand, we were inclined, say, to reject inferences like a = b; therefore, □ a = b—then modal statements involving 'a' would not carry information about what *a* was like in other worlds, and 'a' would be non-rigid. Now we can always ask further questions, like: Why reason in accord with Leibniz's Law? But at some point the answer has to be something along the lines of: Because that's the most useful way to do things, why else?

To summarize: singular and general terms are rigid since directly referential/ directly expressive. They don't contribute semantic values relative to worlds of evaluation, they contribute semantic values simpliciter.

6. Everything Else

So. Names and general terms are all rigid, if Kripke and I are right regarding our respective doctrines. It is left to us to ask whether other simple expressions, such as quantifiers, conjunctions, adjectives, adverbs, complementizers, etc. are rigid, and whether complex expressions, like DP's and CP's, NegP's and DegP's are rigid as well.

With respect to the truth-functional connectives, here is what McGinn (1982) says:

> It appears obvious that they [i.e. truth functional connectives] are rigid and that their rigidity is *de jure*: in evaluating sentences formed from connectives with respect to possible circumstances we need only consider the semantic value the connective has with respect to the actual world - the corresponding truth function is a constituent of the expressed proposition. This is because the semantical rules directly assign those truth functions to the corresponding connectives as fixing

> their contributions to truth conditions; there is no semantic indirection in the selection of negation, say, as that which is relevant to evaluating 'not p' in a world. Again (permitting ourselves some grammatical licence) we can apply the Kripkean intuitive test: can we find a true reading for, e.g., 'Negation might not have been negation'? Clearly not. [pp. 103-104]

I myself am somewhat wary of the "intuitive test" here: first, because the word 'negation' is not the word 'not'; second, because there's some reason to think even Kripke didn't endorse Kripke's intuitive test (see King (2001): pp. 320-321); and third, because if two expressions are non-rigid in the same way (i.e. each refers to what the other refers to in each world, but they change referent from world to world) then the intuitive test delivers the wrong result. Nevertheless, if we assume that 'not' and 'negation' have the same semantic value as one another, and that Kripke's intuitive test works (even if Kripke never endorsed it), then the test provides evidence that 'not' is rigid, and similar tests (under similar assumptions) could be used to show that the quantifiers are rigid. For example, if 'all' and 'universality' have the same semantic value, then the fact that 'universality might not have been universality' has no true reading should be evidence that 'all' is rigid too.

A second line of evidence suggested in the quote by McGinn is to simply ask ourselves whether we consider anything other than the actual semantic value of a simple expression, when evaluating a sentence in which it occurs at other worlds. For instance, when we evaluate 'P and Q' at some non-actual world w, it seems that the sentence is true iff P is true at w *and* Q is true at w. 'And' makes the same contribution relative to arbitrary w that it does relative to the actual world. This line of evidence is easier to pursue than the nominalize-and-Kripke-test strategy considered in the last paragraph. And it seems to straightforwardly extend to all grammatical categories, such as verbs,

adjectives, and adverbs. For instance, for 'John runs' to be true relative to the actual world, John has to run in the actual world. And so for any world: 'John runs' is true relative to arbitrary w iff John *runs* in w. 'Runs' does not denote one activity here, and a different one elsewhere; it expresses the same semantic value; it is a rigid expresser.

A third line of evidence concerns our judgments of the truth and falsity of certain sentences containing modal operators. Suppose for a second that 'not' does not rigidly express a function from propositions to their negations. Then, relative to some worlds, 'not-P' fails to express the negation of P. So it should turn out that ◊ (P and not-P) is *true*. But it isn't. So 'not' is rigid[29]. Similarly, if 'all' sometimes fails to express universal quantification, ◊ (some F's are G, but all F's are not-G) should turn out true (on the further assumption that 'some,' 'but,' 'and,' and 'not' are rigid), but again, it doesn't. Using an actuality operator can reveal quite a bit here. For example it seems false that John might have run, while not doing what is *actually* sufficient for running; and false that John might have run quickly, relative to some standard of quickness, while not doing what is *actually* sufficient for running quickly, relative to that standard. This seems to suggest that 'run' and 'quickly' both persistently express their semantic contents.

In light of such evidence, McGinn says, "we may hazard the generalisation that if an expression is semantically primitive then it has its semantic value *de jure* rigidly (p. 105)." This seems like the right generalization to make.

---

[29] I realize that this argument is not a decisive one, and that it also begs the question. It's not decisive for the following reason. Suppose relative to all worlds except w, 'not-P' expresses the negation of P, but relative to w, it expresses Q, which is logically independent of P. If at w, P is true and Q is false, then 'P and not-P' will turn out false relative to w (because Q is false), and false relative to every other world (because it expresses a contradiction). The argument also begs the question because it assumes that the connective 'and' is rigid, and it is the rigidity of such connectives that is at issue. Nevertheless, the argument is strongly suggestive.

What then of the case of complex expressions? What we say in answer to this question largely depends on how we want to treat the semantics of complex expressions. Consider for example a definite description, 'the F.' If its semantic value, relative to a world w, is the object O that uniquely satisfies F in w (if there is such an O), then clearly, many definite descriptions are non-rigid: 'the inventor of bifocals' has Benjamin Franklin as its semantic value in the actual world, and Descartes as its value relative to some world where Descartes uniquely invented bifocals. If, on the other hand, 'the F' has as its semantic value an intension—for example, a function from a world w to the unique individual O who has F at w, if there is one—then our prior considerations tell us that definite descriptions must *trivially* be rigid: as a direct consequence of the definition of 'intension,' relative to each world, each definite description has the same intension. Finally, if 'the F' expresses a property of properties, namely, the property a property G has when there is a unique F that is G, then it will be a non-trivial question whether 'the F' is rigid, depending on whether 'the F' expresses this same property of properties relative to each possible world.

The case of sentences parallels the case of definite descriptions. If they express truth-values, many are not rigid; if they express intensions (functions from worlds to truth-values), all are trivially rigid; and if they express propositions (where these are not taken to be intensions), the question can go either way. I should point out that on at least one theory of propositions, if simple expressions are rigid, then complex ones are too. This is the 'structured proposition' approach. If the sentence [$_{S}$ *John* [$_{VP}$ *runs*]] has as its semantic value a structured proposition <John <runs>> (where 'John' and 'runs' are the semantic values of '*John*' and '*runs*,' respectively, relative to the context of evaluation),

then if '*John*' and '*runs*' contribute the same semantic value relative to each possible world, [$_S$ *John* [$_{VP}$ *runs*]] will contribute the same semantic value relative to each possible world as well. So if the structured propositionalists are right, and McGinn's hazarded generalization is also right, then rigidity 'percolates up' from the simple to the complex.

For at least some complex expressions, it will be hard to deny their rigidity. For instance, consider 'female fox'. By assumption (that is, following our earlier arguments) 'vixen' is rigid. Hence, if 'female fox' is non-rigid, the sentence 'it might have been the case that vixens weren't female foxes' must be true; but it strikes me that it isn't. The reason it's difficult to tell whether definite descriptions and sentences are rigid is that there's room for debate over their semantic value: objects or second-level functions; truth-values or propositions. But it's hard to deny that 'female fox,' say, expresses the property of being a female fox—that is, the property of being a vixen. And it's telling that when there is no room for debate, complex expressions are clearly rigid.

To conclude, there is a fair amount of evidence in support of the claim that all simple expressions are rigid, that is, that they have the same semantic value relative to each possible world. The evidence is mixed with regard to complex expressions, to a large extent because there is no consensus regarding what the semantic values of complex expressions are supposed to be. And at this point, we must leave the question.

7. Conclusion

For an expression to be rigid is for it to have the same semantic value relative to each possible world. Names are rigid because their semantic value is the object they refer to,

and they refer to the same object, relative to each possible world. In this chapter, I have defended the view that general terms are rigid, because their semantic value is the property they express, and they express the same property, relative to each possible world.

In all, we considered four candidate notions of rigidity, corresponding to four candidate semantic values for general terms: extensions, intensions, satisfiers[30], and properties. We rejected the thesis that rigidity was rigid extension, because none of the paradigm cases of rigid general terms came out rigid, according to the thesis. We rejected the thesis that rigidity was rigid intending, because it is trivially true that all general terms are rigid intenders. And we rejected the thesis that rigidity was rigid application for two reasons: first, because like rigid extension, none of the paradigm cases of rigid general terms come out rigid, according to the thesis; second, even if the paradigm cases were rigid, according to the thesis, nevertheless it is powerless to explain why 'water is H_2O' is necessary, if true.

Thus I endorsed the thesis that rigidity for general terms was rigid expression: a general term G is a rigid expresser iff G expresses the same property relative to every possible world; or, alternatively: iff *if* for any P, G expresses P relative to the actual world, *then* G expresses P relative to any other possible world. It was argued that on this account all general terms are rigid, but that this was a substantive—not a trivial—fact. Furthermore, I argued that my view could explain why 'water is H_2O' and 'tigers are animals' are necessary, in exactly the same way that Kripke's notion of rigid designation allowed him to explain why 'Hesperus is Phosphorus' is necessary.

[30] That is, things that satisfy the general term, things that it *applies* to.

There are several upshots to these conclusions. First, Leibniz's Law for properties is that $(X)(Y)(X \equiv Y \rightarrow (\Phi \leftrightarrow \Psi))$, where X and Y range over properties, and Φ and Ψ differ at most in that open occurrences of X in Φ are replaced with Y in Ψ. In simpler terms, numerically identical properties have numerically identical intensions. Intuitively, every substitution-instance of the law involving a general term is true: if the property of being a groundhog is numerically identical to the property of being a woodchuck, then necessarily, all and only groundhogs are woodchucks; and if the property of being a water closet is numerically identical to the property of being a bathroom, then necessarily, all and only water closets are bathrooms. As outlined in Section 4, the doctrine that rigidity for general terms is rigid expression, combined with the doctrine that all general terms are rigid expressers, delivers the result that every instance of Leibniz's Law for properties is true, thus confirming intuition.

Second, we are now in a position to unify and simplify the semantics of general terms. If it is not just natural kind terms, but all general terms that are rigid, then all general terms may be given a unified semantic treatment. This is the Uniformity Principle in action: what is true for natural kind terms is true more generally for all general terms. Further, if rigidity for general terms is rigid expression, i.e. if rigid general terms express the same property, relative to each possible world, then we may simplify our semantic clauses by stripping them of superfluous world-parameters. We need no longer say "'dog' expresses, relative to some world w, f(w)" for some specified function f from worlds to contents, but rather may more simply say "'dog' expresses (simpliciter) the property of being a dog."

Chapter 3

Against Compositionality (As a Metasemantic Thesis)

I. Introduction

In virtue of what does a complex expression mean what it does, rather than something else, or nothing at all? One potential answer to this question is *compositionality*. A complex expression means what it does in virtue of its parts and their combinatory structure. The meanings of the parts and the way in which they are combined *metaphysically determines* the meaning of the whole. To get a sense of what I mean when I say 'metaphysically determines,' the claim I have in mind is that complex expressions get their meanings from their parts and how they're combined in much the same way that simple expressions get their meanings from the causal or informational relations they bear to objects and properties in the world. That is, I'm reading 'compositionality' here as a metasemantic thesis.

This isn't the traditional understanding of compositionality, and for good reason. Traditionally, compositionality is interpreted along these lines: a language is compositional iff there is a way of computing the meanings of complex expressions from the meanings of their parts and how they're combined[31]. This formulation is entirely neutral with respect to metasemantics; it says nothing about what it is in virtue of which the complex expressions have their meaning.

[31] To be unfair, "traditionally" compositionality is construed in a much narrower fashion, so that for example Kamp's Discourse Representation Theory and Tarski's treatment of bound variables are "non-compositional." I'm utterly unmoved by *this* tradition and so I ignore it henceforth.

One way of seeing this is that some proponents of compositionality so-understood endorse reverse compositionality: that one can compute the meanings of the constituents of complex expressions from the meanings of the complex expressions themselves (cite Fodor). Clearly 'in virtue of' can't go both ways: it can't be that wholes have their meanings in virtue of the meanings of their parts and that parts have their meanings in virtue of the wholes they compose. Compositionality so-understood is simply a thesis about the relation between the meanings of parts and the meanings of wholes, namely, that it's computable. This tells us nothing about the metaphysics, any more than the fact that one can compute the prime decomposition of 27 tells us something about the metaphysical dependence of 3 on 27 or vice versa.

Thus, let's have names for two different claims. On the one hand, we have what I'll call compositionality as a metasemantic thesis (or CMT) and on the other, compositionality as a computability thesis (CCT). The purpose of this paper is to argue that CMT is false *for the language of thought*; I intend to be utterly indifferent as to the truth of CCT, and for the moment I intend to suspend judgment on whether CMT is true or false for natural language.

But why, you might ask, dispute a claim (CMT) that philosophers don't typically make? Well, just because they don't make it doesn't mean they don't think it's true. Take for instance Dretske's (1981) metasemantic account for mentalese expressions. Here, expression E has as its content property P because tokenings of E are correlated with P in the learning environment. No one would seriously entertain such an account for complex expressions: there is no learning environment in which the expression NIXON IS PURPLE is correlated with Nixon's being purple. Dretske's metasemantic account for simple

expressions is only plausible given the assumption that a different metasemantic account is to be had for complex expressions, presumably something like CMT. NIXON IS PURPLE means that Nixon is purple because NIXON is correlated with Nixon in the learning environment, IS PURPLE is correlated with being purple in the learning environment, and the meaning of the whole is metaphysically determined by the meanings of the parts and their mode of combination. That's the standard story, that's the story I aim to dispute.

But why, you might ask, ought we to go about disputing CMT? If so many metasemanticists of differing stripes, from Dretske to Millikan to Fodor, have simply assumed its truth, what gives us cause to pause? Well, this is my own hobby horse and you're welcome to feel differently about it: uniformity. I'm inclined to think that something along the following lines is true: if meaning is anything, it's one thing. I don't mean by 'meaning' the things meant, those differ as wildly as you can imagine (literally). Rather, I mean the metasemantic relation between an expression and the thing it means. If simple expression E means P because it bears R to P, then simple expression E' should mean P' because it bears R to P', and complex expression E'' should mean P'' because it bears R to P'', and so on.

I used to think that descriptivism with respect to proper names preserved at least some of this uniformity. For suppose names are descriptive. Surely definite descriptions are descriptive. So names and definite descriptions are metasemantically uniform: a name gets its semantic value in exactly the same way that a definite description gets its semantic value, by description. But that's not right[32]. Descriptivism with respect to proper names says that a name N refers to an object O because N is associated with a

[32] It was Jeff King who pointed this out to me.

description that O uniquely satisfies. That's not how definite descriptions work, at least on the standard story: a definite description D has its referent determined by the referents of its parts and their syntactic mode of combination. Definite descriptions aren't descriptive.

But this sort of reasoning cuts both ways. Definite descriptions—again, according to the standard story—don't have their meanings determined causally either. So if uniformity is worth having, then no one has something that's worth having. It seems a shame on either construal of the quantifiers (though much worse on one of them).

So is uniformity important? I think so, but it's sort of a hunch. If meaning exists (that is, if semantic eliminativism of the Stich/ Field sort is false), it's probably important. And its importance, if it has any, is probably tied to the role it plays in cognition and communication. Having true generalizations about meaning's important role would allow us to explain and predict things we wouldn't otherwise be able to explain and predict, or so I'm inclined to believe. So that's the first part of the hunch, that meaning exists and is important due to its role in cognition and communication. The second part of the hunch is that disparate, disjunctive things tend not to be important, and generalizations involving such things tend not to explain or predict much at all. It's unified things that do all the explaining. I take it these hunches are of the sort that drives a lot of philosophy: we assume that if there's any notion of knowledge worth having, there's exactly one such notion; if there's any notion of the good worth having, there's exactly one such notion; any notion of cause, etc. etc.

Of course, this isn't my argument against CMT. As I said, you're welcome to take uniformity or leave it. I'd like to think you'd take it if you could get it, and so part of my

argument will be that you can get it. But there's another motivation for denying CMT that some may find more compelling.

On one view of linguistic semantics the goal is this. The theorist (semanticist) is tasked with assigning to each unambiguous natural language expression some formal representation (its "meaning") and to each syntactic mode of combination some effective procedure for composing the "meanings" of the terms so combined into a new "meaning," the "meaning" of the whole. The criterion of adequacy for such an assignment is that the representations and algorithms so assigned are strongly equivalent[33] to the representations and algorithms implemented by the mind of a speaker successfully interpreting expressions of the language. The outcome of an adequate semantic theory then is the theorist's ability to predict and explain speakers' judgments of intuitive derivability relations (and as a special case, judgments of intuitive interderivability, that is, judgments of synonymy). I have spoken with a handful of working semanticists who assure me this is what they do and perceive that it's what they're paid to do. Let's assume this is what linguistic semantics is.

An important corollary of this assumption is that linguistic semantics isn't about *semantics* at all, at least if the latter is construed as the study of which things in the world are the contents of which natural language expressions. 'Hesperus is bright' has as its content something like the proposition that Hesperus is bright, and this proposition entails and is entailed by the proposition that Phosphorus is bright. But this fact is not a fact a semanticist endeavors to explain. Rather, she wants to explain why speakers judge that 'Phosphorus is bright' does not follow from 'Hesperus is bright' and she does this accordingly by assigning 'Hesperus' and 'Phosphorus' distinct mental representations. In

[33] In the sense of Pylyshyn (1984).

this way, semanticists are a species of cognitive psychologists: the goal is to model a certain aspect of natural language processing that is downstream of speech segmentation and parsing and upstream of pragmatic processing.

But suppose we do ask about the external-world contents of our expressions. An optimist might think that the world obliges, that ontology recapitulates philology. For example, if the linguistic semanticist assigns a discourse referent representation to ‘a man’ in the sentence ‘a man walks in’ then there’s some thing in the world, a discourse referent, that is the sort of thing it could be true of to say ‘walks in’ (see Cumming (2007) for an ontology of such things). Or, for another example, if the linguistic semanticist assigns a proposition representation to ‘how to swim’ in the sentence ‘A man knows how to swim’ then there’s a proposition *how to swim* that the discourse referent referred to by ‘a man’ is related to in virtue of which it’s true to say ‘A man knows how to swim’ (see Stanley & Williamson (2001)). And so on and so forth.

Although I’m an inveterate optimist, I’m inclined to be pessimistic in this instance. Ontology doesn’t recapitulate philology any more than ontogeny recapitulates phylogeny. And this is the norm. The representations and algorithms posited by cognitive psychologists are almost never a good guide to events and the dynamical laws that govern them. Witness: folk physics stinks. You can’t get physics from folk physics, and you can’t get ontology from folk ontology. Our world might well consist of a universal wave function that occupies an infinite dimensional phase space—that is, it might be true that ‘our world consists of a universal wave function occupying an infinite dimensional phase space’—without the universal wave function being a discourse referent and without the phase space being a discourse referent. I mean, they could well be a *wave function* and a

phase space, after all. And it might well be that to know how to tie your shoes involves being able to execute a program that eventuates in your having tied shoes, in such a way that propositions never enter the picture.

Now think of what this means for the compositionality of mentalese, under the provisional assumption that the representations and algorithms posited by linguistic semanticists are strongly equivalent to the representations and algorithms of mentalese. And by 'compositionality' I don't mean what the linguistic semanticist means—that is, the fact that there is an effective procedure for mapping representations syntactically combined into composite representations—but rather CMT: the composition of the external-world contents of mentalese representations with one another. It might well be that what the semantics delivers is not what the metaphysics delivers, and that what the semantics delivers can't be mapped one-to-one into what the metaphysics delivers, that is, it might well be that CMT is false.

Here's an example of what I'm thinking about. There are infinitely many expressions of the form 'has a mass of 5 kg'. Now suppose that we work out the metaphysics and it turns out that there are infinitely many *primitive, unstructured* mass properties. There's a property of having a mass of 5 kg, but there's not a part of that property that's the "mass" part and a part that's the "5" part and a part that's the "kg" part. In fact, suppose there's absolutely no referent for "mass" (nothing *is* mass), and no referent for "5", and no referent for "kg". But there are *primitive, unstructured* properties that objects tend to have when the scale says they weigh 5 kg.

CMT defenders should be very afraid of such metaphysical possibilities. After all, this example exactly parallels Scheffler's view of indirect discourse that Davidson (1984) argues against as follows:

> Scheffler suggests we analyse 'Tonkin said that snow is white' as 'Tonkin spoke a that-snow-is-white utterance'… the expression 'that-snow-is-white' is to be treated as a unitary predicate (of utterances or inscriptions)… the syntax is clear enough… But there is no hint as to how the meaning of a predicate depends on its structure. Failing a theory, we must view each new predicate as a semantical primitive. Given their syntax, though (put any sentence after 'that' and spice with hyphens), it is obvious there are infinitely many such predicates, so languages with no more structure than Scheffler allows are, on my account, unlearnable. [p. 12]

The same could be said for predicates like 'having a mass of 5 kg'. The syntax is clear enough, but there is no hint as to how the meaning of such a predicate depends on its structure: its parts have no meanings. Given their structure, though (stick a numeral between 'having a mass of' and 'kg'), it is obvious there are infinitely many such predicates. If what I've described is genuinely a metaphysical possibility, then we must countenance either the possibility that we've learned an unlearnable language (English) or that CMT is false.

Here, then, is a preview of the picture I'll have on offer. Syntactically simple expressions have their referents determined by the causal/ informational/ lawful relations they bear to objects and properties. Whatever that story is, it's precisely the same story for syntactically complex expressions. The intentionality of complex expressions is primitive and not derived, in particular, it's not derived from the referents of their simple constituents and their mode of combination. In this way, the metasemantics treats complex expressions as though they had no internal structure.

2. Preliminaries: Compositionality as a Metasemantic Thesis

I'm afraid I can't say too much in the way of clarifying exactly what CMT is supposed to amount to. To make it clear why, let's consider one way you might think of understanding CMT, which is really a way that you shouldn't understand it.

Szabó (2000) has argued that compositionality should be understood as a strong supervenience thesis, in particular, the thesis that the meaning properties of complex expressions strongly supervene on their "constitution properties"—properties of the form: "*having some constituents with such and such meanings combined in such-and-such way*" (p. 495). In formal terms:

> For all possible [human, first] languages *L*, for any meaning property *M* and any complex expression *e* in *L*, if *e* has *M* in *L*, then there is a constitution property *C* such that *e* has *C* in *L*, and for any possible [human, first] language *L* if any complex expression *e* in *L* has *C* in *L* then *e* has *M* in *L*. [p. 499]

I do want to read CMT in such a way that it entails the strong supervenience thesis SST, but SST itself, as I read it, is too weak for the reverse entailment to hold.

First, consider what SST does and does not rule out. Suppose a language L contains an expression X meaning *gray* and an expression Y meaning *elephant*, and that it contains the same adjective-noun syntactic combination rule as English—that is, [X Y] in L has the exact same syntax as [gray elephant] in English. Then SST requires that [X Y] means in L what [gray elephant] means in English, namely, *gray elephant*. This seems right: compositionality *is* violated when two complex expressions with the same constitution properties have distinct meanings.

However, SST does not rule out the following case. The case is not to be construed counterfactually, but rather as a fictional supposition about how the actual world is. Here's what we want to keep the same:

> 1. All the lexical items in fictional English have the meanings they have outside this fictional supposition.
>
> 2. The syntax of fictional English is as it is outside this fictional supposition.
>
> 3. Complex expressions that are synonymous in English as it is outside this fictional supposition are also synonymous in fictional English.

But here's what's different. In fictional English, 'gray elephant' means *black swan*, 'gray horse' means *blue chicken*, 'red donkey' means *fried eggs*, and similarly for all adjective-noun combinations: their meanings have absolutely nothing to do with the meanings of their parts, in an entirely unsystematic fashion.

This scenario isn't possible, according to SST, because English [gray elephant] means *gray elephant* and thus any possible language where 'gray' means *gray* and 'elephant' means *elephant* is such that [gray elephant] means *gray elephant* in it. But SST does not rule out the possibility that this fictional scenario is actual, since by supposition no two expressions of fictional English with the same constitution properties have distinct meaning properties (see (3) above)[34].

Now you'd be right not to be worried that English is like this, because it isn't. The point I want to make, however, is that CMT, as I want to understand it, should entail that compositional languages aren't like this. CMT says that the meanings of complex

[34] I'm fudging a little here. We'd have to add to the supposition that expressions in languages other than English that have the same constitution properties as English expressions were synonymous with those English expressions. And we'd also have to suppose that this were so even in other languages that were possible but not actual in the fictional scenario—but the scenario itself plus SST entails that, so I'm assuming that's already part of the scenario. The finer details here, however, are really beside the point I'm making.

expressions depend upon their constitution properties, and 'depend' here is something stronger than mere strong supervenience. Complex expressions, says CMT, *get* their meanings from their constitution properties, their meanings don't merely necessarily co-vary with their constitution properties.

What's more, I don't think one can just trump up SST to get the right result. If we wanted, we could read CMT as SST + CCT. This would rule out the case described, because presumably its unsystematic character prevents a recursive treatment. But still something is missing. The provable truths of arithmetic strongly supervene on the physical structure of the world, because no two worlds that are physical duplicates have distinct provable truths of arithmetic. And the provable truths of arithmetic are computable from the physical structure of the world, since they're computable without it, and arithmetic is monotonic. But they don't *depend* on the physical structure of the world. They don't hold *in virtue of* that structure. CMT is about the metaphysical dependence of meaning properties on constitution properties, not a thesis about necessary co-variation or computability.

I apologize that I can't say more about CMT. I hesitate to suggest that the murky character of the dependence at issue might be a further reason for rejecting CMT, since that smacks a little too much of substituting my ignorance for an argument. I'll say at least this. Whatever CMT amounts to, it *entails* SST + CCT, though as I've said the entailment isn't mutual.

3. Preliminaries: Mentalese vs. Natural Language

I say here that I'm only concerned with denying CMT for the language of thought, not natural language. I suppose I should also say why I have the concerns I have.

On what I take to be the standard picture, complex natural language expressions have a metasemantics twice removed from simple mentalese expressions. One remove is already familiar: on the standard story, complex natural language expressions derive their semantic values from the semantic values of their parts and how those parts are syntactically combined. If this remove were the only remove, I'd have no compunction with making my argument against CMT for complex natural language expressions as well.

There is, however, a second remove. On the standard story, simple natural language expressions derive their semantic values not from the causal, informational, or lawful relations they bear to things in the world, but rather inherit them from the mental expressions they are conventionally associated with (if I'm not being clear, I just mean broadly Gricean, early-Schiffer-style Intention Based Semantics). Thus arguing against CMT for natural language would not net me any uniformity points whatsoever, so long as I wasn't also prepared to argue against Intention Based Semantics. For all I'd wind up with is that complex natural language expressions get their meanings in one way (not compositionally, but rather inherited directly from the complex natural language expressions they are conventionally associated with) and complex mental expressions get their meaning in an entirely different way (namely, causally).

And am I willing to argue against Intention Based Semantics? Perhaps someday, but the issues are tricky and would unnecessarily complicate what here is already a complicated exposition. The difficulty with running a causal theory directly on natural

language expressions is quite severe. Suppose me to be an unscrupulous buyer and seller of wares. When you bring me something worth $150, I am disposed to say "It's worth $100!" and when I show you something worth $50, I am disposed to say "It's worth $100!" In no case is the actual value of an object the cause of my saying "It's worth $100!" No causal theory I know of is capable of handling such cases.

So there you have it. Natural language is not my target here, though natural language will loom large in what follows, as it is wont to do.

4. How Could CMT be False?

I imagine the biggest worry about the thesis that CMT is false is: how could it be false? What possible story can one tell on which complex expressions have their semantic value primitively, and not in virtue of what their parts mean? And why has no one proposed such a story before, if possible it be?

I want to start with simple expressions and ask the same question: how could they get their semantic values primitively? One line of thought (the only serious one I'll be considering in this paper) runs as follows. There are on the one hand simple mental expressions (henceforth: concepts) like COW. Concepts are individuated syntactically, so that the concept COW and the concept X are identical iff one can infer, with no premises whatsoever, that COW = X (i.e. the inference rules of mentalese license the inference from the null set of premises to COW = X). On the other hand there are properties in the world, like the property of being a cow, where properties are ways that things (in general) can be. (I say "in general" here because I mean that properties are ways that some things

can be, not ways that any thing can be; a cow couldn't be a horse, but being a horse is a way some things could be—for instance, horses could be [and are] that way.)

According to the broad class of metasemantic views I want to consider, a concept C has as its referent the property P iff the property of tokening C stands in some appropriate causal, informational, or lawful connection to the property of being P. For example, P's are the most natural class of things that cause tokenings of C; tokenings of C have the function of carrying information about things being P; or there's a law that P's cause tokenings of C's, and any other law that Q's ($Q \neq P$) cause tokenings of C's depends on the former law but not vice versa; or etc. etc.

Let's, for the moment, take the latter as being our paradigm case (this is Fodor's Asymmetric Dependence Theory—ADT). How does the theory explain why COW refers to the property of being a cow, and not some other property or no property at all? Well, first, things that instantiate the property of being a cow cause tokenings of COW in virtue of their instantiating that property. (Compare this: things that instantiate the property of being called 'LeBron James' don't cause basketball victories in virtue of their instantiating that property—if I legally changed my name to 'LeBron James' I'd be no more likely to win in the NBA than if I changed it to 'Woody Allen'. However, if I became a cow, I would be much more likely to cause tokenings of COW.)

But this can't be the whole story. First, there are problems with error: horses on dark nights may cause me to token COW in virtue of their being horses on dark nights. After all, horses on dark nights reliably look like cows on dark nights. Second, there are problems with robustness: my lack of a cow may cause me to token COW (in the form of a desire that I have a cow) in virtue of it's being a cow-lack (it's certainly not causing the

desire in virtue of it's being a lack-on-Tuesday). Yet COW doesn't refer to the property of being a cow or being a horse on a dark night or being a cow-lack or…

So the second part of the story is the asymmetric dependence of these laws. Something's being a horse on a dark night causes you to token COW because you mistake it for being a cow. If you didn't *take* cows for being cows, you couldn't *mis*take horses on a dark night for being cows. So the fact that you can mistake horses for cows depends upon the fact that you can take cows for cows—or, rather: the law that connects horse-on-a-dark-night instantiations to COW tokenings depends upon the law that connects cow instantiations to COW tokenings—but notably, not vice versa. The dependence is asymmetric: if you could somehow never mistake horses for cows, you could nevertheless take cows for cows. The latter law is independent of the former.

The same goes for robustness. Tokening CHICKEN causes you to token COW, because you've associated chickens with cows. But if cows didn't cause you to think COW (and say instead that skyscrapers did), why would you associate CHICKEN with COW? It's because cows make you think COW that COW and CHICKEN are associated in your mind—but notably not vice versa. The law that connects CHICKEN tokenings to COW tokenings depends on the law that connects cow instantiations to COW tokenings. If you weren't to think of cows when you thought CHICKEN, it wouldn't follow that you wouldn't think of cows when you thought COW. The latter law is independent of the former.

The same story, I claim, can be told about complex expressions like BROWN COW. Clearly, things that instantiate the property of being a brown cow can cause you to think BROWN COW. By supposition, their instantiation of the property of being brown

causes you to think BROWN; and their instantiation of the property of being a cow causes you to think COW; and you can, in general, figure out that the brown thing is the cow thing: in situations where you think BROWN(x) and COW(y) where as a point of fact x = y, you can be caused to think x = y and thus infer BROWN(x) & COW(x). So there is a law connecting a thing's being a brown cow to your tokenings of BROWN COW.

You'll make errors of course. Sometimes a thing's being a red cow, a brown horse, or a red horse, will upon occasion cause you to token BROWN COW. But unless you took brown cows to be brown cows (i.e. unless instantiations of brown cowhood lawfully caused BROWN COW), you wouldn't mistake red cows and brown horses for brown cows. The law that connects red cow instantiations to BROWN COW depends on the one that connects brown cow instantiations to BROWN COW, for much the same reason that the law connecting tokenings of red to BROWN depends on the law connecting tokenings of brown to BROWN.

Robustness is to be handled in much the same way. You might well associate brown cows with red barns, and thus it may be lawful that your thinking BROWN COW causes you to think RED BARN. But were brown cows not to cause BROWN COWs, that is, were brown cows not to dispose you to think of them, then presumably they wouldn't dispose you to think of red barns either. The asymmetric dependence theory delivers the results we want, in the same manner, whether it's applied to simple or complex expressions. And this is how CMT could be false: the meaning of complex expressions just might run through the direct lawful relations such expressions bear to extramental objects, without regard to what the parts of the expressions in question mean. It's the goal of the rest of this chapter to make this idea plausible, if not even palatable.

5. First Argument *against* CMT: The ABC Argument

Suppose you are investigating an intelligent agent and discover the following. When the agent perceives that a thing is small for its size, or otherwise has evidence that something is small for its size, it tokens a symbol X; when the agent perceives that a thing is orange, or otherwise has evidence that a thing is orange, it tokens Y; and when the agent perceives that a thing is a ferret, or otherwise has evidence that a thing is a ferret, it tokens Z. You investigate the matter quite thoroughly, and are fairly confident in your assessment that X means *small*, Y means *orange*, and Z means *ferret*.

Furthermore, suppose you observe that the agent is disposed to infer as follows: from X, Y, and Z to XYZ, and from XYZ to X, XYZ to Y, and XYZ to Z. Thus, whenever the agent perceives that a thing is a small orange ferret, it is disposed to token XYZ, and when it tokens XYZ, it is disposed to treat the target of its thought as a small orange ferret. It seems that the only appropriate content to assign to XYZ is *small orange ferret*.

But this situation gives rise to two distinct metasemantic possibilities. On the one hand, it might be the case that XYZ means *small orange ferret* because X means *small*, Y means *orange*, Z means *ferret*, and the meaning of the whole depends upon the meanings of the parts. XYZ inherits its meaning from X, Y, and Z. Alternatively, however, it could be that XYZ means *small orange ferret* because it is small orange ferrets that (appropriately) cause the agent to token XYZ. Nothing in the case as described can tell between these two possibilities. We need a different case.

Suppose then that the intelligent agent you are investigating has the following peculiarity. When the agent perceives that a thing is large for its size, or otherwise has evidence that something is large for its size, it tokens a symbol A; when the agent perceives that a thing is brown, or otherwise has evidence that a thing is brown, it tokens B; and when the agent perceives that a thing is a cow, or otherwise has evidence that a thing is a cow, it tokens C. You investigate the matter quite thoroughly, and are fairly confident in your assessment that A means *large*, B means *brown*, and C means *cow*.

That's not the peculiar part. You further observe that the agent is disposed to infer as follows: from A, B, and C to AB, and from AB to A, AB to B, and AB to C. What content are we to assign to the complex expression AB? It would seem that the appropriate answer is *large brown cow*. When the agent perceives that a thing is a large brown cow, it is disposed to token AB, and when it tokens AB, it is disposed to treat the target of its thought as a large brown cow.

If this is right, then CMT can't be true. If the meanings of the parts of every complex expression determine the meaning of the complex expression, then AB should mean *large brown thing*, not *large brown cow*.

This argument says nothing about CCT. Here, we were supposing to have found such an agent, but it doesn't follow from that that there are such agents, and especially not that we are such agents. It could turn out that as a point of fact and for any actual agent, one can always compute the meaning of a complex mental expression from the meanings of its parts. What the ABC argument is supposed to show, however, is that the meaning of the complex expression doesn't depend on the meanings of the parts, it depends on the things that (appropriately) cause its tokening. In all cases (XYZ, AB) the

correct content assignment to a complex expression tracks the things that (appropriately) cause its tokening, but only in some (perhaps nonactual) cases is the correct content assignment to a complex expression some function of the meanings of its parts.

One tempting line of response to the ABC argument goes something like this: the argument falsely assumes that AB is a complex symbol built out of A and B; rather, it is a simple symbol, something like an idiom for mentalese. But I don't think one should give in to the temptation. Ask yourself this: *why* should we think that AB is simple? If your answer was "because it doesn't get its meaning from the meanings of its parts and their mode of combination," I think you should reconsider.

What you really don't want to say is that AB's being simple consists in its being non-compositional. For CMT says that whether a thing is compositional (inherits its meanings from its parts) depends on whether it's simple or complex, and it can't also be the case that whether a thing is simple or complex depends on whether it's compositional. Dependence is antisymmetric.

Now you might want to say that AB's being simple doesn't *consist in* its being non-compositional, but its being non-compositional is nevertheless *evidence for* its simplicity. But I fail to see what gives you the right to make that claim. Admittedly, the standard tests for idiomhood look to see whether an expression is compositional (cite). But they do so because they assume, methodologically as it were, that all the complex expressions of the language under investigation are compositional. Thus finding a seemingly non-compositional expression is extremely reliable evidence that that expression is simple and not complex (i.e. is an idiom). But notice that in a debate with me, you don't get to help yourself to the methodological assumption that all complex

expressions are compositional, at least not if you're committed to not begging the question.

To reiterate, the point of the ABC argument is this. From the standpoint of the compositionalist, there is *no reason at all* to expect that a complex expression meaning, say, *brown cow* is caused by the presence of brown cows, or causes behavior that would be successful were there a brown cow around. This is because complex expressions, according to the compositionalist, get their meanings from their parts, not their causal relations with objects and properties in the world. They could have any or no causal relations with objects and properties in the world and still mean what they do because of what their parts mean. So it's perfectly consistent for the compositionalist to hold that AB means *large brown thing-in-general* even though it's caused solely by large brown cows. Thus there shouldn't be the slightest modicum of a reason for the compositionalist to think AB was an idiom or whatever—no part of her theory requires or even recommends doing so. In fact, her theory tells her not to do so. Why then do we feel so pulled to think that AB means large brown cow? Not to put to fine a point on it, but let me suggest it's because CMT is false.

Carlotta Pavese (p.c.) has suggested to me another way one might want to resist the ABC argument. One of the premises of the argument is that it's established that A means *large*, B means *brown*, and C means *cow*. But shouldn't we simply deny this claim, especially in light of the fact that AB is caused by large brown cows, not large brown things in general? Wouldn't a cognitive psychologist in this precise situation actually give a more nuanced meaning for A, or for B?

Well, yes, maybe a cognitive psychologist would do that, but again, this isn't what her theory tells her to do. I mean, let's suppose that we're Dretske circa 1981, and that simple expressions mean what they do because they were correlated with their contents in the learning environment (LE) and that complex expressions mean what they do because they inherit their meanings compositionally from their parts. Suppose furthermore that A is correlated with largeness in LE, B is correlated with brownness, and C with cowness. This suffices to determine what AB means. What causes AB and what AB causes is strictly speaking irrelevant to what AB means on the theory in question. My point, however, is that it sure as heck *seems* to be relevant.

6. Second Argument *against* CMT: Causally Isolated Things

One of the big worries about causal theories in general is that certain things seem to stand outside the causal order—things like conjunction and disjunction, and numbers and numerical functions. The worry is that three plausible theses seem to be in conflict with one another:

> 1. Metasemantic uniformity for simple expressions: if simple expression E refers to object O because E bears R to O, then for any simple expression E', E' refers to O' iff E' bears R to O'.
>
> 2. Causal isolation: disjunction is causally isolated. Disjunction never causes anything, much less causes our tokenings of 'or'.
>
> 3. 'Fred is a cat or a dog' is true iff Fred is a dog or a cat.

For suppose that some causal theory is true, and that 'dog' has its content because of the causal relations it bears to the property of being a dog. Then by (1), 'or' must have its content be X, where X is what 'or' bears the appropriate causal relation to. Yet by (2),

‘or’ can’t bear any causal relation to disjunction, so either ‘or’ is meaningless or it means something other than disjunction. Finally, if ‘or’ doesn’t mean disjunction, then how can (3) be true? For if ‘Fred’ refers to Fred, ‘is a cat’ refers to the property of being a cat, ‘is a dog’ refers to the property of being a dog, ‘or’ refers to nothing or something other than disjunction, and the referent of the whole depends on the referents of the parts, then how can ‘Fred is a cat or a dog’ refer to the disjunctive state of affairs: either Fred is a cat or Fred is a dog?

Notice however that the paradox is reached only by assuming compositionality. On the other hand, if we reject compositionality and hold that ‘is a cat or is a dog’ can refer to the property of being a cat or a dog, even if ‘or’ is referentless, the paradox is avoided. It might be thought that if disjunction is causally isolated, so must all disjunctive properties be (like the property of being a dog or a cat). But I don’t know of anyone who seriously entertains this. After all, were it true, there wouldn’t be a disjunction problem—we’d never have to worry that ‘cow’ meant *cow or horse-on-a-dark-night*, because the latter property would, being disjunctive and thus by supposition causally isolated, immediately be ruled out of contention as a referent in some causal theory. Furthermore, I’m inclined to think that ‘disjunctive property’ is a category mistake: the property of being a cat or a dog is numerically identical to the property of not being a non-cat and a non-dog—properties don’t have logical structure (though they do bear logical relations to one another).

We can extend the strategy to other expressions as well. It’s plausible to think that the number 7 is causally isolated, yet that ‘there are seven cats’ refers to the state of affairs of there being 7 cats. This is only a problem for me if the property of being seven

cats is causally isolated. And presumably it isn't: the apartment manager threw me out of my apartment *because* I had 7 cats and the limit was 6.

I don't want to commit myself to the claim that disjunction, say, or the number 7 actually are causally isolated. I'm inclined to think, for example, that a group of cats can instantiate sevenness, and their so instantiating this property can cause various things, among them my thinking SEVEN. But I tend to get funny looks when I argue this in public, and it's hard to tell a similar story for OR. So no more: 7's putative causal isolation may be a problem for other causal theorists who want to uphold metasemantic uniformity for simple expressions, but it's not a problem for me, since I can always simply deny compositionality (i.e. deny CCT; I'm committed to always denying CMT).

7. Frege Cases

Since my claim is that complex expressions acquire their referents directly via causal or informational connections they bear to things, and this immediately entails that sentences acquire their referents via bearing those same relations to things, it's appropriate to ask what, according to me, are the things that are referred to by sentences?

The first thing I'll note is what, according to me, they're not. The referents of sentences are not structured propositions, and this follows simply and straightforwardly from the fact that structured propositions have more structure than the things I bear causal, informational, or lawful relations to. If water causes X, then H_2O causes X; if Y carries information about water, then Y carries information about H_2O; and if the property of being water is lawfully connected with the property of being Z, then the

property of being H_2O is lawfully connected with the property of being Z. A proponent of structured contents may well want to distinguish the content of 'water' and the content of 'H_2O,' and she may well want to treat the latter as finer grained in virtue of its greater structure. But then her notion of content cannot be my notion of referent, because the referent of 'water' (namely, water) and the referent of 'H_2O' (namely, H_2O) play exactly the same role in causation, information, and lawful connection, the central components of any causal theory.

The same is true, *mutatis mutandis*, for the referents of sentences. That the board was lighter than water caused it to float; it's floating carried the information that it was lighter than water; and the board's being lighter than water is lawfully connected to its floating. None of these facts, however, are changed by substituting 'H_2O' for any or all of the occurrences of 'water' in the preceding statement, or for that matter, by substituting 'that water was heavier than the board' for 'that the board was lighter than water.' The "propositiony" things that play causal, information, or lawful roles are much less fine grained that structured propositions[35].

So what are they? I will take them to be a sort of property. If you'll recall, for me, a property is a way that a thing could be. Since worlds are things, ways that worlds could be are properties, and I wish to identify the referents of sentences with ways that worlds could be. To introduce some terminology, I'll call a way that a world could be a *state of affairs*, and I'll say that when a world w could be some way P and w *is* P, then the state of affairs P *obtains* in w.

[35] Master Argument that sentences don't have structured propositions as their contents: I can see what you mean; but I can't see propositions; so propositions aren't what you mean.

Let me be clear that I do not take myself to be committed to the claim that states of affairs are no more fine grained than sets of possible worlds. I'm already committed to the thesis that states of affairs aren't identical to sets of possible worlds. Sets of worlds are objects (in particular, they're sets) and objects aren't properties, that is, they're not ways that things can be. Two distinct states of affairs may obtain in all and only the same worlds without being identical. I won't spend much time exploiting this fact, so I'll limit myself to a reasonably unilluminating example: the state of affairs of there being something trilateral is not equivalent to the state of affairs of there being something triangular.

So I'm not a structured propositionalist or a sense theorist or anything like that, and given the rest of my views, I can't be. But that doesn't mean my view doesn't share some of the distinct advantages of these various fine-grained accounts of content. In particular, one of the biggest concerns for the causal metasemantic theorist is no concern of mine: Frege cases.

Frege's Puzzle comes in two flavors, one mentalistic and the other attributivistic. My view can easily handle both. The mentalistic version goes something like this: How can it be that agent A believes that Hesperus is bright while at the same time not believing that Phosphorus is bright, given that 'Hesperus is bright' and 'Phosphorus is bright' refer to the same state of affairs? Here, I take it, it's legitimate to plead syntax. To believe that Hesperus is bright is to have in one's belief box the mentalese sentence HESPERUS IS BRIGHT and to believe that Phosphorus is bright is to have in one's belief box the sentence PHOSPHORUS IS BRIGHT. It's easy to see then how you could believe that Hesperus is bright but not believe the same of Phosphorus. Perhaps you lack the

PHOSPHORUS concept; perhaps you haven't computed the entailment from HESPERUS IS BRIGHT and HESPERUS = PHOSPHORUS to PHOSPHORUS IS BRIGHT; or perhaps you can't compute the entailment, because HESPERUS = PHOSPHORUS is not in your belief box.

The attributivistic version of Frege's Puzzle is markedly different. It goes like this: How can it be that "Agent A believes that Hesperus is bright" is true while at the same time "Agent A believes that Phosphorus is bright" is false, given that 'that Hesperus is bright' and 'that Phosphorus is bright' refer to the same state of affairs? And my answer is: it could be this way if compositionality were violated. That is, it could be that in both sentences 'Agent A believes' has the same referent, and that 'that Hesperus is bright' and 'that Phosphorus is bright' have the same referent, but the referent of the whole does not depend on the referents of the parts, so "Agent A believes that Hesperus is bright" and "Agent A believes that Phosphorus is bright" refer to distinct states of affairs. For instance, the first might causally/ informationally/ lawfully depend on A's having HESPERUS IS BRIGHT in her belief box and the second might causally/ informationally/ lawfully depend on A's having PHOSPHORUS IS BRIGHT in her belief box, and as we've just seen, it's possible for these two states of affairs to come apart.

It is only the person who is (a) a referentialist (b) not a structured propositionalist (or a Fregean sense theorist or whatever) and (c) committed to either CMT, SST, or CCT that has a problem with Frege's Puzzle. In particular, no one numerically identical to me has this problem.

Of course, in saying this I do not mean to commit myself to the claim that compositionality is violated in propositional attitude ascriptions, only the claim that *if*

"Agent A believes that Hesperus is bright" is true and "Agent A believes that Phosphorus is bright" is false, *then* compositionality is violated (i.e. CCT is false). If however these two statements have identical contents, as some philosophers have maintained, it is open to me to say that the referents of each are computable from the referents of their parts (though of course not dependent on the referents of their parts).

8. Where We've Been and Where We're Going

Here is a summary of the case so far for denying CMT, and affirming a direct causal theory for complex expressions:

> 1. **Global Uniformity**: There's something to be said for everything having its meaning determined in the same way. If CMT is right, then the way that complex expressions get their meanings is different from the way simple ones do. So there's something to be said for denying CMT, and adopting the metasemantic theory for simple expressions as the metasemantic theory for complex expressions.
>
> 2. **Local Uniformity**: To the extent that there's something to be said for everything having its meaning determined in the same way, there's even more to be said for all simple expressions having their meanings determined in the same way. But it looks impossible to tell a causal or informational story for OR or SEVEN without denying CMT. So there's even more to be said for denying CMT, and adopting the metasemantic theory for simple expressions as the metasemantic theory for complex expressions.
>
> 3. **Intuitive Correctness**: In cases where what CMT urges and what the direct causal theory for complex expressions urges come apart, intuitively the direct causal theory is right. In a case where a complex expression AB is caused appropriately by large brown cows, and causes behavior that would be successful in the presence of large brown cows, we are intuitively inclined to say that AB means large brown cow, *regardless* of what its parts mean.
>
> 4. **Ontological Neutrality**: Alas, ontology rarely recapitulates philology, or mentalese philology (if I may). If CMT is right then how we refer to a property P puts constraints on our metaphysics. For instance, if P is the semantic value of the complex expression XYZ, then there must be an aspect of the world that is X's

> semantic value, an aspect that is Y's semantic value, and an aspect that is Z's semantic value, and P has to be determined (in the metasemantic way) by these aspects and the structure of the expression XYZ. But in many cases it seems likely that while P is ontologically respectable, putative values for X, Y, and Z are not (e.g. in the measure phrases case, described in Section 1).

(Note the similarity between (4) and the causal isolation argument in (2). The arguments aren't the same, because (2) is concerned with the causal commitments of CMT and (4) with its metaphysical commitments. But they are species of the same genus.)

> 5. **Free Pass on Frege Cases**: Frege cases are notoriously tough, especially for the causal theorist. But they only arise when we assume CMT or something like it (SST or CCT, for example). This is not to say that there are not possible compositional treatments of Frege cases available to the causal theorist. But a direct causal theory for complex expressions has no problem handling the cases. (*This is not to say it has no other problems!*)

That then, is the positive case for my view. In what follows, I'll consider what I take to be the principal objections to the direct causal theory of complex expressions. As a preview of what's to come, here they are:

> 1. **Computability**: According to me, CMT entails CCT, that is, that the meanings of complex expressions can be computed from their constitution properties. But the direct causal theory is at best independent of CCT, and at worst incompatible with it (for example, in causal isolation cases and Frege cases). Yet CCT appears on its face to be true, at least for natural language. And what goes for natural language may well go for the language of thought. If so, that's evidence that CMT is right and I'm wrong.
>
> 2. **Learnability and Understandability**: We can produce and understand a potential infinitude of complex natural language expressions. In much the same fashion, we can produce and understand[36] a potential infinitude of complex mentalese expressions. Many have thought that compositionality and compositionality alone can explain these abilities. If so, that's evidence that CMT is right and I'm wrong.
>
> 3. **Systematicity**: Language L is systematic iff For all expressions X, Y of L of the same category, and all modes of syntactic combination C in L: for any Z, C(X, Z) is a well-formed expression of L iff C(Y, Z) is. Systematicity is often thought

[36] Which seems to be all one in this case, that is, producing = tokening = thinking = understanding. It's not as though your thoughts come to you, and then you must figure out what they mean. And if they did, we'd have an infinite regress on our hands. And our hands are too small for infinite things to be on them.

to bear some close connection to compositionality, though the precise relation is a matter of some debate and of some obscurity. But any argument that systematicity entails compositionality is an argument against me, if mentalese is indeed systematic.

4. **Problem Cases**: There seem to be cases where an agent A tokens a complex mentalese sentence S, and S refers to some state of affairs P, but P never, wouldn't, and couldn't cause S to token in A. Thus the direct causal metasemantic theory for complex expressions seems to be unable to assign the correct referents to such sentences. However, the compositionalist has no such problem with the cases: provided the parts have meanings, the wholes do. So that's evidence that CMT is right and I'm wrong.

Some of these problems, I'll argue, aren't problems at all, or at least not problems for me. Others are, and a truly honest account, such as I wish to have, will have to bite some not particularly tasty bullets as a result. In the conclusion, I'll try to weigh the balance of evidence as impartially as I can.

9. Computability

There's a rather successful research program in linguistics that's in the business of detailing how the meanings of complex natural language expressions can be computed. It's somewhat of a methodological presupposition that this can be done for all complex natural language expressions, and there's very little that's been discovered that makes it look like it can't. Although the principal target of this paper is complex mental expressions rather than complex natural language expressions, I think it's important to address the relation between compositionality and computability in linguistic theory, if I'm to be above board.

Recall the picture of linguistic semantics from the introduction. There, I characterized linguistic semanticists as a species of cognitive psychologists. Their goal, I

maintained, was to specify, or assume from the work of syntacticians, the data structures output by the mind's parser (i.e. syntactic representations) and an algorithm that took, as one of its inputs, such data structures and mapped them to formal representations, the "meanings" of the expressions, which I shall here, for want of a better term call 'semantic representations.'[37] The methodological assumption that there is such an algorithm strikes me as utterly unimpeachable: if semantic representations can't be computed from other representations upstream of them, how in the world does the mind get from the one to the other? If I were a nativist, I'd say that Mysterianism just isn't in my blood; if I were an empiricist, I'd say that Mama didn't raise no Mysterian. However that turns out, linguistic semantics seems on solid footing to me.

But something's missing in the argument from this computability thesis to CMT. After all, CMT entails the thesis that people can compute the meanings of expressions, but not vice versa. One man's inference to the best explanation is another man's affirming the consequent. And I'm the other man.

Why doesn't the thesis that people can compute the semantic values of linguistic expressions entail that such expressions have a compositional semantics? Simply put, the algorithms taking you from syntactic representations as inputs to semantic representations as outputs *might require other inputs*, up to and perhaps including *every other representation in your mind*. It's a methodological presupposition, or if you like a transcendentally deduced a priori constraint, that the semantic values of at least that

[37] Though this is somewhat of an oxymoron, or a redundancy, depending on how you see it. *No* semantic value is a representation (except in the case of the semantic values of, say, quoted expressions standing for representations) and *every* representation has a semantic value (except in the case of defective representations). But by 'semantic representation' I really mean 'representation assigned by a linguistic semanticist in the project so outlined,' rather than 'representation that is a semantic value' or 'representation that has a semantic value.'

subset of expressions of natural language we're capable of understanding can be computed, but that doesn't begin to buy you the assumption that they can be computed from only the syntactic representations and the entries in the lexicon.

Just to take a concrete example, Kamp's Discourse Representation Theory (DRT) (1981) is non-compositional in precisely this sense. The semantic value of some linguistic expression E in DRT often depends upon the semantic values already computed for expressions occurring earlier in the discourse and thus ipso facto does not depend solely on the structure of E and the features the lexicon assigns to the parts of E. I could, if I didn't have the reader's best interests in mind, multiply such examples indefinitely. An Optimality Theoretic approach to semantic computation might well hold that two complex expressions E1 and E2 in two languages L1 and L2, which have the same structure, and the same exact entries in the lexicon for each of their parts, might nevertheless differ in meaning (that is, differ in the semantic representation assigned to them by the 'semantic mechanism') when the languages in question differed in their relative rankings of faithfulness and markedness constraints. And the point here is not that such theories might not be replaceable by compositional ones—there are indeed compositional treatments of the phenomena handled by DRT (see Groenendijk & Stokhof (1991))—the point is that methodological considerations or transcendental deductions can't tell you whether the mind, in its actual workings, goes Kamp's way or instead goes Dutch.

The inference from the unimpeachable claim that speakers can compute the semantic representations of linguistic expressions they understand to the claim that they must do it with solely the inputs from the syntactic mechanism and the lexicon is, I think,

of dubious merit. Where does that leave us with respect to CMT? Well, I think the inference from the unimpeachable claim to CMT is of doubly dubious merit. I made this point in fn. 6, but I was careful there to disguise it so as not to tip my hand too early.

The point is this. Although when we understand a complex linguistic expression, we compute its semantic values, we don't compute the semantic values of mentalese expressions when we understand them. To understand a mentalese expression is all one with thinking it, and thinking it is all one with tokening it. And things *couldn't be otherwise*. If we had to compute its semantic value, we'd have to do so in a metamentalese, which presumably to be understood would have to have the semantic values of its representations computed in a metametamentalese, and so on, right up the infinite hierarchy. There's an excellent a priori argument that we compute the semantic values of linguistic expressions we understand, because we pair them with corresponding semantic representations, and this isn't done by magic. But there's an equally excellent a priori argument that we don't compute the semantic values of mentalese expressions, because this would involve pairing each of them with a higher-level semantic representation, ad infinitum, which can *only* be done by magic.

This is not to say that the semantic values of mentalese expressions aren't computable, or even that they aren't computable from their constitution properties. All I'm arguing is that we needn't compute their semantic values to understand them, and so without a further argument we can't conclude that their semantic values are in fact computable. Furthermore, even if their semantic values are computable, which has yet to be shown, it doesn't follow that they're computable from their constitution properties. So as successful as compositional natural language semantics is, I don't think it has the first

thing to tell us about whether metasemantics is compositional, that is, whether CMT holds.

So we don't compute the meanings of mentalese expressions when we understand them. But it might be urged that they are nevertheless computable, and even computable from their constitution properties, because, well, they seem that way. To get the meaning of BROWN COW you take the meaning of BROWN and the meaning of COW and conjoin them, or apply one to the other, or whatever. This isn't part of the process of understanding, but it may be part of the process of assigning contents to the representational states of agents we're investigating. And wouldn't this be good inductive evidence that CMT is true after all?

Maybe, but I'm inclined to doubt the premise—the premise about how mentalese seems. Just because I hit the 'caps lock' button on my keyboard when I want to represent mentalese representations doesn't mean that mentalese is a sort of caps-lock English, where there's a homomorphism from the syntactic structures of English into those of mentalese. Conscious thought and planning is often accompanied by the perception as of English sentences spoken aloud[38], but we know that mentalese isn't English because we think before we can speak, and we think things we cannot speak. A psychologist might have something to say about the structure of thought, but our phenomenology doesn't.

Again, to summarize: there's no good argument from computability to compositionality. And even if there were, there's no good argument that the meanings of

[38] This has always puzzled me. If you don't mind some speculation, here goes (if you do mind, this footnote is not for you): maybe we are incapable of consciously experiencing our thoughts, as opposed to our perceptions, which we *are* capable of consciously experiencing, when we attend to them. And maybe, to get around the limitation, evolution has co-opted the mechanism that produces the speech stream from our thoughts to the service of producing an 'inner' speech stream, run through the auditory circuits (and thus perceived auditorily), so that we can at least be conscious of the natural language expressions of our thoughts, even when the thoughts themselves remain inaccessible. This is all a bit of a fairy story without any empirical evidence herewith adduced, but I did warn you in advance.

mentalese expressions are computable from their constitution properties. So however the project of linguistic semantics stands, CMT stands or falls independently of it.

10. Learnability and Understandability

The argument from learnability to compositionality goes something like this. It's impossible for a finite mind in a finite amount of time to learn infinitely many things. If complex expressions had their meanings primitively, and not in virtue of the meanings of their parts, then the meaning of each complex expression would have to be learned separately. But there are infinitely many complex expressions (or at least we have the capacity to understand any of infinitely many such expressions). Therefore, complex expressions don't have primitively determined meanings.

To my mind, this line of reasoning isn't particularly compelling when applied to the language of thought. Here's my reply. From the perspective of a causal metasemantic theory, having an expression E that refers to an object O is a matter of E bearing a certain causal relation to O. The agent need not represent something like E REFERS TO O, and thus there is no infinite set of such representations that the agent must learn. All that's required is that O would cause E in the appropriate circumstances. Unless there's some reason to think that a potential infinitude of states of affairs would not cause a potential infinitude of mentalese sentences, then there's no reason to think that complex expressions can't have their referents determined primitively.

The case can be made even stronger. Suppose that the premise of the objection is correct: that we really do have an ability to produce and understand an unbounded set of

representations. It follows (by &-elimination) that we have an ability to produce that set, and thus that each member of that set *could* be caused. Thus, a causal theory can assign contents directly to each member of that set—perhaps highly implausible contents, perhaps not[39]. So no general consideration can lead us to suppose that directly assigning contents to complex expressions can't work, although of course a demonstration that the contents assigned in particular cases were incorrect could lead us to suppose that.

Not to sound like a broken record, but this point is implicit in what's gone before. It *can't* be that understanding a mentalese expression E requires representing something like E REFERS TO O. Because then O is a mentalese expression, and understanding it would require representing O REFERS TO O*, and so on, ad infinitum. Unless the mind is infinite, this is impossible, and that the mind is finite is exactly what my hypothetical objector is urging. Thus understanding mentalese expressions must involve bearing a mind-external relation to their contents, not a mind-internal relation to some further representation. And general considerations like the fact that the mind is finite won't be sufficient to put substantive constraints on this relation, like the constraint that the content-conferring mind-external relation a complex expression bears to its semantic value metaphysically depends upon the content-conferring mind-external relation its simple constituents bear to their semantic values. That could be so, but a learnability/ understandability argument doesn't get you in the environs.

11. Systematicity

[39] I should say *probably* not. After all, the referents assigned will be the things that under normal conditions would cause you to token those representations (or whatever your favorite causal story is), and the lesson of the ABC argument was that these are *very* plausible candidates for the semantic values of those representations.

Here's a formal characterization of systematicity:

> Language L is systematic iff For all expressions X, Y of L of the same category, and all modes of syntactic combination C in L: for any Z, C(X, Z) is a well-formed expression of L iff C(Y, Z) is.

If you don't like that characterization, fine. I just want you to have the flavor of the idea, and I'd prefer not to use any of the informal characterizations running around the literature, as they're not as flavorful.

The putative[40] systematicity of the language of thought has been used as a premise in two distinct arguments for two distinct claims, presented here:

> **Claim about Mental Architecture**: The language of thought is symbolic in the same sort of way that natural language, logic, and artificial computer languages are. It does not contain "distributed" representations of the PDP/ connectionist/ artificial neural network type. (Fodor & Pylyshyn, 1998[41])

> **Claim about Semantic Structure**: The meanings of complex expressions literally contain the meanings of their parts. (Sometimes this is called "reverse compositionality," but I've been informed that there's a limit on how many different things can be called 'compositionality' in one paper, and that I've already exceeded it.) (Fodor & Lepore, 2001[42])

I'm going to assume that if there's an argument against the direct assignment of content to complex expressions, it's not the putative systematicity of mentalese, as that at best is merely an observation, but rather it's some important property of mentalese of

---

[40] I say 'putative' because I mean it: see what is, in my mind, a devastating critique of the claim that natural language is systematic and a compelling critique of the claim that mentalese is so: Johnson (2004).

[41] "There is… a straightforward (and quite traditional) argument from the systematicity of language capacity to the conclusion that sentences must have syntactic and semantic structure: If you assume that sentences are constructed out of words and phrases, and that many different sequences of words can be phrases of the same type, the very fact that one formula is a sentence of the language will often imply that other formulas must be too: in effect, systematicity follows from the postulation of constituent structure."

[42] "What we'll call 'reverse' compositionality… assum[es] that the meanings of constituent expressions supervene on the meanings of their complex hosts… the explanation [for the supervenience] is obvious: the meaning of 'dogs bark' supervenes on the meanings of 'dogs' and 'bark' because the meanings of 'dogs' and 'bark' are parts of the meaning of 'dogs bark'; and *the meaning of 'dogs' and 'bark' supervene on the meaning of 'dogs bark' for exactly the same reason.*"

which its systematicity is a symptom—a property like its syntactic structure (architecture) or the semantic structure of the contents of its expressions. That is, I assume that if there's an argument against my view in the vicinity here, it starts with one of the two Claims above.

So consider the Claim about Mental Architecture. If it's in conflict with my view, then my view is in conflict with one of my deeply held commitments, that the mind is a computer. But I don't think the Claim really is in conflict with my view. Let's assume for simplicity's sake that the language of thought is first order predicate calculus with identity. Then my view is that complex expressions like $(x)(Fx \rightarrow Gx)$ get their contents directly, via what (appropriately) causes them, and not indirectly, via their constitution properties. Roughly: everything's an idiom. But everyone agrees that idioms are possible, even ones with complex syntactic structure. So there can't be any conflict between the view that complex expressions are syntactically structured and that they're idiomatic. At least not any logical conflict.

Now maybe there's conflict when you assume that all the expressions in the language are understandable. Then you might think infinitely many idioms are too many idioms. But that was the argument we covered in the previous section, so I'll leave it there.

If I can endorse the Claim about Mental Architecture—and I can because I do and nothing prevents me from doing so—I get systematicity for free, for exactly the reason Fodor & Pylyshyn point out: "systematicity follows from the postulation of constituent structure." So it looks like systematicity can't really be a consideration in favor of rejecting my view.

What about the Claim about Semantic Structure? Well, I'm committed to denying it. Recall that for me, 'or' has no meaning, and so its meaning is certainly not a part of the meaning of 'John is a cat or a dog.' At least not a part of it in the spirit in which that claim is intended. Further, I think that 'John believes Hesperus is bright' might well fail to contain the meaning of 'Hesperus' as a part. It might mean the state of affairs: John has the mentalese sentence HESPERUS IS BRIGHT in his belief-box. And Hesperus just isn't a part of that state of affairs.

Admittedly, this doesn't put me in shabby company. All of contemporary semantics that falls under the broad heading of Montague Grammar—including Stalnaker and Lewis, the DPLers, and the online update/ direct compositionality crowd—rejects the Claim about Semantic Structure. This is because, on all these accounts there exist what Johnson (2006) calls "conflating contexts"—the input to semantic composition is not recoverable from the output. So, for instance, just as 3 and 4 aren't recoverable from 7 given only the knowledge that you got to 7 by addition (because you could get there by any of the pairs <0,7>, <1,6>, <2,5>, <3,4>, etc.), so too the functions λx. dog(x) and λP. P(a) aren't recoverable from dog(a), given only the knowledge that you got to dog(a) by function application (because you could get there by a and λx. dog(x) just as well, as well as by infinitely many other instances of function application). Thus the meaning of the whole doesn't contain the meanings of the parts, contra the Claim, nor do the meanings of the parts supervene on the meaning of the whole.

Johnson argues that, with respect to the linguistic data, everyone must countenance conflating contexts. His principal example comes from basic data involving the progressive morpheme in English. "Telic" verbs are those verbs that are semantically

specified as having an endpoint. This is clearest in the "creation verbs" like 'build' and 'bake' which are such that if you built a house, a house existed (there was a completion of the building process), and if you will bake a cake, a cake will exist (there will be a completion of the baking process). There are verbs that aren't verbs of creation which are telic, such as transitive 'run': if you ran a mile, there is a completion of the event (your having traversed a mile by running), even though nothing comes into existence by your actions in this case.

The progressive morpheme in English maps telic verbs to atelic ones and atelic verbs to atelic ones. For instance 'Mary was building a house' does not entail 'Mary built a house' because the latter but not the former is semantically specified for an endpoint: you can have been building a house even when your activity never eventuates in a house, whereas you cannot have built a house without such an eventuation (similarly, you don't get to go to the finals just because you were winning the game; you have to have won it). However, atelic verbs are unaffected by progressivization: 'I watched a cat' and 'I was watching a cat' are mutually entailing.

It therefore follows that a verb V in the context "progressive marker + V" is in a conflating context. The input verb (e.g. 'build' or 'watch') is specified for telicity or atelicity. The output verb (e.g. 'building' or 'watching') is always atelic. Thus some aspect of the meaning of the input verb is unrecoverable from the output. The Claim about Semantic Structure must be false, because the semantics of [be [build ing]] does not fix the semantics of [build] (it doesn't tell you whether [build] is telic or atelic).

I'm convinced by Johnson's reasoning, and I think this is something structured propositionalists and neo-Fregeans should be worried about. But in this context I really

don't care to harp on the point. The real point is: it may well be that natural language, though systematic, nevertheless doesn't validate the Claim about Semantic Structure. As a result, there's even less reason to suppose mentalese validates the Claim, because there's even less reason to suppose we know lots of substantive things about mentalese. As such, the fact that I have to deny the claim strikes me as not at all expensive, in the grand scheme of things.

In conclusion, there's no obvious contradiction between the putative systematicity of the language of thought and my view. Furthermore, neither of the Claims that are supposed to follow from systematicity are of much threat to me: one is apparently compatible with everything I say, and the other is apparently incompatible with what most people say, and with what the evidence seems to say. I conclude, not incorrectly in my mind, that it's time to move on.

12. Some Other Worries

I've claimed that we can assign states of affairs as referents directly to sentences, without considering and sometimes without regard to the referents of the sentences' parts. I've claimed that we can do this by determining which states of affairs (appropriately) cause which sentences. That seems terribly contentious though. Why not think that there are cases where an agent A tokens S, S refers to P, but P never, wouldn't, and couldn't cause S to token in A? Cases like these:

Case 1: A doesn't know which observations are confirmatory of S.

Suppose on her biology exam, Amanda is given a picture of cell and is asked: "How many mitochondria are present in this cell?" Amanda can't tell a mitochondrion from a ribosome, so she just guesses, and writes "There are three mitochondria." Here it certainly seems that though what she wrote means that there are three mitochondria, and refers to the state of affairs of there being three mitochondria, this state of affairs, being unrecognizable to Amanda, cannot ever lawfully cause her to token what she wrote (or what she was thinking when she wrote it). So there seems to be a conflict here with what my theory says and what the facts on the ground are.

I'm inclined to think that this isn't much of a problem for me. After all, Amanda can't tell a mitochondrion from a ribosome. So why does 'mitochondrion,' when she says it, refer to the property of being a mitochondrion? There are lots of stories out there—one of them is "deference to experts." Since experts' tokenings of 'mitochondrion' bear the appropriate causal relation to mitochondria, and Amanda defers to those experts, her tokenings of 'mitochondrion' mean what theirs mean.

If that story is fine, why isn't this one? Amanda can't tell the state of affairs in which there are three mitochondria from the state of affairs in which there are three ribosomes. So why does 'there are three mitochondria,' when she says it, refer to the state of affairs of in which there are three mitochondria? Answer: deference to experts. Since experts' tokenings of 'there are three mitochondria' bear the appropriate causal relation to the relevant state of affairs, and Amanda defers to those experts, her tokenings of 'there are three mitochondria' mean what theirs mean.

I'm not a huge fan of deferency stories. They're a tad non-uniform: Amanda's term 'mitochondrion' gets its referent from the causal relations *experts* bear to mitochondria, whereas her term 'shoe' gets its referent from the causal relations *she* bears to shoes. But there are other stories. In the testing situation, Amanda is barred from looking in her textbook. This is because if she could, the fact that mitochondria look like such-and-so (which caused the textbook illustrator to illustrate them looking like such-and-so) could cause her to think that mitochondria looked like such-and-so. Then the fact that three mitochondria are depicted in the test question could cause her to token 'there are three mitochondria.' That is, there is a lawful route from the state of affairs to the mentalese sentence, and those who have studied as well as those who have cheated have found it.

There are of course tougher cases. Sometimes there aren't experts, not even possible ones. Consider this statement: "What we think of as our universe is a tiny part of a larger world, and from that vantage point appears much like a point particle appears to us. And that world is contained similarly by a bigger world, and so on and so on ad infinitum. Furthermore, things that appear to us as point particles are actually quite complex worlds, with inhabitants as complex and varied as in our universe. And what appear to the inhabitants of those worlds as point particles are also complex and varied worlds, and so on and so on ad infinitum." That certainly seems for all the world like a contentful statement. But how could its being true ever bring it about that we thought it was true? Here there are no experts to exploit or defer to: no one in our speech community can or even could evaluate that statement for its truth with any authority.

If I may, I'll bite this bullet: you simply cannot think things for which there could not be confirmation conditions. I suppose that makes me somewhat of a verificationist (though, mind you, I'm *not* identifying contents with confirmation conditions). I'll have more to say in the next section about how I want to think about the bullets I've bitten. At the moment, I just want to identify the ones that need biting.

Case 2: Statements about the future

The future is problematic for me for the following reason: causation runs in only one direction, from the past to the future. Perhaps if time travel happens, there's some backwards causation, but not a lot, and not enough. Most future states of affairs do not and cannot cause us to token sentences that intuitively refer to them. This is pretty bad for the causal theorist.

Information often replaces causation in "causal" theories for precisely this reason. The stormcloud on the horizon carries the information that it will rain; this information causes me to token 'it will rain'; so we might well say that sentences refer not to states of affairs but to the information that certain states of affairs obtain[43]. This hardly solves any of our problems: most future events are literally unpredictable even if predictable-in-principle: there's no causal pathway from the information that these future events will occur to our thinking of the events—we don't have access to the relevant information. Other future events are unpredictable-in-principle (like the outcome of an indeterministic

[43] Alternatively, we could take statements about the future to refer to the present state of affairs that such-and-such will happen, if we admit such states of affairs into our ontology. The same problems will arise.

process): there's no information *to* have access to. Yet we think of and plan for such contingencies.

Bringing laws into the picture can help in the following way. Consider some unpredictable but predictable-in-principle state of affairs P. Were P predictable, the information that P could then cause us to think that P will obtain. Furthermore, on the assumption that time travel is possible, nothing is in principle unpredictable: we could always receive word of it through a wormhole. So there is a law connecting the information that P with our thinking that P, it's just that we don't see it in play when the information that P is inaccessible. There's nothing ad hoc in appealing to such laws—we do it in the case of simple expressions as well. When I don't have access to the information that cows are around (say because I'm blind and there's no one here to describe the countryside), cows don't cause COW. But we still say that COW refers to cowhood because when we do have access to this information, cows *do* cause COW.

It should be heartening to hear that simple expressions recapitulate the problem for complex expressions. Consider the word 'future.' Nothing possessing the property of futurity causes tokenings of 'future.' And if we had a single word meaning 'unpredictable future occurrence' (we don't but we could), then the information (if it existed) that some unpredictable future occurrence would occur would never cause us to token it. So the causal theorist for simple expressions is going to have to say roughly what I've been saying, or otherwise solve my problems for me. The boat had better be big, because we're all in it.

Case 3: Although S is true, A is constitutionally incapable of believing S, regardless of the evidence presented to her.

If there's no causal process from the state of affairs P, or from the information that P, to A's tokenings of S, it can't be that S refers to P, at least on any causal or informational account of reference. Nevertheless, there seem to be clear cases where S does refer to P, but no such process exists.

For example, consider George, who under no circumstance is inclined to believe that his son Fred is gay (that is, George won't token S, S = 'Fred is gay', no matter what evidence he has). For instance, Fred says to George "I'm gay"; George is present at Fred's legal marriage to another man; George believes that Fred's husband is gay, etc. etc. Nevertheless, George does not think that Fred is gay, and does not token FRED IS GAY in response to any of this evidence, nor would he, given more evidence.

Someone who accepts CMT can easily account for the phenomenon. George's concept FRED refers to Fred because it is appropriately caused by Fred; his concept BE GAY refers to the property of being gay because it is appropriately caused by that property; and the whole FRED IS GAY refers to Fred's being gay because the meaning of the whole is determined by the meanings of the parts and their syntactic mode of combination. However, I cannot say this, because I deny CMT.

The first thing to say is that the problem is not just a problem for me, but also for the causal theorist of simple expressions. Consider George's concept GAY. Why does it refer to the property of being gay, rather than the property of being a gay person who is not Fred? After all, it's the second connection that is lawful (gay-but-not-Fred → GAY),

not the first. And tokenings of GAY in Fred's mind carry information about gays who aren't Fred, not about gays sans phrase.

This point can be sharpened. Suppose for a moment that God exists. I don't really care which God, so pick your favorite. Bridgette is a militant atheist. Extremely militant. Under no circumstance will God cause Bridgette to token GOD. No amount of miracles, divine revelations, sublime experiences, or what-have-you will lead Bridgette to token GOD. Yet Bridgette believes GOD DIDN'T SPECIALLY CREATE MAN. So she does token GOD, it's just that there's no causal pathway from God to GOD. The causal theorist must deny that her concept GOD refers to God and that her belief that GOD DIDN'T SPECIALLY CREATE MAN is true iff God didn't specially create man. Why then should it be so hard to deny that George's belief FRED IS GAY is true iff Fred is gay?

I think the intuitiveness of the claim that George's belief FRED IS GAY refers to the state of affairs of Fred's being gay arises from the extreme difficulty of actually imagining a circumstance in which someone was so unresponsive to reasons that no amount of evidence could move them.

But supposing such were the case, I'm not sure denying that their mental states had content is so absurd. Imagine a creature with X in its desire-box. Nothing in principle could cause the creature to token X in its belief box. The creature, as a matter of principle, will never take its desire to have been fulfilled, no matter the state of the world and no matter the evidence the creature has about the state of the world. Now try assigning a content to X. I don't know on what basis you would choose.

And it doesn't strike me that the case of complex expressions is that much different. In that case, of course there is a basis for you to choose a content. You choose

the content that can be derived from the meanings of the parts and their syntactic modes of combination. I can see why this is a reasonable strategy: since typically the referents of wholes can be computed from the referents of their parts, it's not absurd that we'd simply use this strategy to assign contents to recalcitrant representations. But I also don't see what would stop us from saying these representations had no content, just as in the case of X. If there's no theoretical virtue to taking the compositionalist route (and there isn't, because he still has to deal with Bridgette's GOD either way), then I'm inclined to bite the bullet. George's fear FRED IS GAY had no content, because no appropriate causal pathway runs from Fred's being gay to George's tokening FRED IS GAY.

Case 4: P is impossible

Statements about impossible states of affairs seem especially difficult to handle from a causal viewpoint. As Fodor once remarked, "nothing cramps one's causal powers like not existing."[44] Impossible states of affairs don't cause anything; nothing carries information about them; and plausibly, they aren't lawfully related to possible states of affairs (like tokenings of certain sentences). Yet without a doubt certain sentences do refer to impossible states of affairs (if they refer to states of affairs at all). For example, consider the sentences "There exists a round square" and "2 + 2 = 5."

This problem, if it is a problem, is not just a problem for causal theories of complex expressions (such as the one on offer here), but also a problem for causal

[44] Fodor, J. 2008. "Against Darwinism." *Mind & Language*, 23: pp. 42-49. I just like the quote; I don't mean to start a debate on natural selection here.

theories of simple expressions, like those expounded by Dretske, Millikan, Fodor, et al. Fodor (1990) sort of recognizes this:

> [If the asymmetric dependence theory is true] *no* primitive symbol can express a property that is necessarily uninstantiated. (There can't, for example, be a primitive symbol that expresses the property of being a round square… [T]he notion of primitiveness that's at issue here isn't entirely clear. You could, presumably, have a *syntactically* primitive that means *is a round square* so long as it is 'introduced by' a definition. [p. 101]

So Fodor thinks he can avoid trouble by finding a space between primitive symbols and syntactically primitive symbols, though he makes it clear a few lines later that he doesn't know how to do this. I don't see why he doesn't just say that a primitive symbol is one that has its meaning primitively, that is, via the causal relations it bears to objects and properties in the world. So the reason *syntactically* primitive expressions with stipulated impossible content don't count as primitive is that they inherit their content from the parts of the right-hand side of the stipulation, just as syntactically non-primitive expressions inherit their content from their own parts.

I obviously can't say this however, because that's not how I want to handle stipulations and that's not how I want to handle complex expressions. Everything is primitive, in the sense I've suggested above, so it would seem nothing refers to what's impossible. But, I think, if that puts me badly off, Fodor himself is no better off. To reiterate, he says: "[If the asymmetric dependence theory is true] *no* primitive symbol can express a property that is necessarily uninstantiated." But consider the primitive symbol 'impossible.'[45] If it expresses the property of being impossible, then it expresses a property that is necessarily uninstantiated. And if it doesn't, what then does it express?

[45] Please don't tell me that 'impossible' is a complex expression in English. First, it isn't. It isn't 'in-' + 'possible' because English doesn't have the rule np > mp, Latin does (cf. unpredictable); the word was borrowed as a unit into our language. Second, it doesn't matter. Take a simple irrealis mood marker from

As with some of the previous cases, we might advance a little by moving to a causal theory cashed out in terms of laws. For example, some have thought that counterfactuals are grounded in laws. The reason that it's true that were it raining, the streets would be wet is that there's a law, roughly: ceteris paribus, when it rains the streets get wet. Yet there seem to be counterfactuals connecting impossible states of affairs with possible ones: for example, if one could square the circle with just a ruler and compass, Hobbes would be even more respected than he is. It seems equally plausible to suppose that it might be true that if one could square the circle with just a ruler and a compass, I would token "one could square the circle with just a ruler and a compass." And so it might be a law that ceteris paribus, when one can square the circle with just a ruler and a compass, I token "one can square the circle with a ruler and a compass." If one runs one's causal theory via laws (as, say, Fodor does), this might suffice to establish the appropriate meaning constituting relation.

Furthermore, Putnam (1975) has argued that logic itself is an empirical matter. It might well be then that there are possible worlds in which modus ponens gets violated, and the statement "Modus ponens is sometimes invalid" refers to the state of affairs wherein modus ponens is sometimes invalid because were we in such a world, that state of affairs would cause us to make the statement.

This sort of reasoning, however, will not work in general. For suppose that it's true that there are no statements that are false in every possible world, and thus that every state of affairs obtains in some world or another and is therefore available as a referent

---

some other language, and the example still goes through. Finally, there's no reason to suppose that the morphological structure of English words is an accurate reflection of the structure of mentalese. The point here is just that there may well be a simple expression of mentalese meaning "impossible" so the causal theorist, even of simple expressions, had better have a story to tell.

for statements seemingly describing impossible states of affairs. If no statement is false in every world, then there must be some world relative to which the statement S = "there *are* statements that are false in every possible world" is true. Suppose one such world (where S is true) is w. Then in w there is some statement P that is false in every world; so there is some statement that is false in every possible world (by elimination of the vacuous 'in w'); so our initial supposition was false.

It follows that I'm committed to the claim that some sentences that CMT-endorsing theorists can assign content to are such that I can't assign content to them. So be it. The result isn't so bad, after all. For one, I can still say why they seem meaningful. Suppose ONE CAN SQUARE THE CIRCLE WITH A RULER AND COMPASS is one of the things I'm forced to say is referentless. I'm not, however, forced to say that one can't token it; nor that one can't infer it from the perfectly referentful HOBBES SAID ONE CAN SQUARE THE CIRCLE WITH A RULER AND COMPASS; nor that one can infer from it the perfectly referentful COMPASSES EXIST. If meaningful things can lead you to think it, and it can lead you to think meaningful things, it's no wonder it seemed meaningful. But it isn't, not if I'm right.

For another, I'm not in bad company. The Russell of (1910) thought that falsehoods in general were referentless, because sentences, when they referred, referred to facts, and falsehoods weren't facts. But I have it that only necessary falsehoods are referentless, with the garden variety of falsehoods referring to non-actual but possible states of affairs. In this way, I'm quite like Stalnaker—he assigns the null set to necessary falsehoods and I simply don't give them any referent (and this would make a difference if I endorsed compositionality—I'd have to assign referents to disjunctions of truths with

necessary falsehoods—but since I don't, it doesn't). I actually have a leg up on Stalnaker in that I don't have to assign the same thing to all necessary truths.

13. Summary of Cases and Replies

All of the previously considered 'bad' cases had something in common: the impossibility of a state of affairs causing a mentalese sentence that intuitively referred to it. This impossibility had different sources in different cases. For example, the state of affairs itself might be impossible, or it might be possible but impossible to encounter, or it might be possible, possible to encounter, but the agent encountering it might be incapable of tokening the appropriate sentence upon such an encounter.

My replies to these cases all had the same general bi-partite form. First, I pointed out that none of the particular problems in question were unique to me, but each also must be faced by the causal theorist of simple expressions. Impossible things don't cause 'impossible', inconceivable things don't cause 'inconceivable', future things don't cause 'future', etc. The lay of the land then seems to be that exactly one of (1)-(3) must be true in its 'simple' version and exactly one must be true in its 'complex' version:

1. Causal metasemantic theories of simple (complex) expressions are false.

2. A solution to these problems exists and it can be applied to causal metasemantic theories of simple (complex) expressions.

3. No solution to these problems exists, but it is acceptable if causal metasemantic theories of simple (complex) expressions bite the relevant bullets.

Without a compelling reason for thinking the problems arise for different reasons in the simple and complex cases, our default assumption should be that 'simple'-(1) is

true iff 'complex'-(1), and *mutatis mutandis* for (2) and (3). Then all I need is: *not 'simple'-(1)*—a not unreasonable background assumption I suppose many share with me already—and my view is off the ground.

Of course, that argument is only so intellectually satisfying. My second strategy of response was to attempt to defuse some of the problem cases by showing that they weren't as far-reaching as might be supposed at first glance. So, for instance, it might be the case that we can treat (deterministic) future cases in the same way we treat present cases with inaccessible information that is in principle accessible (ribosomes don't cause RIBOSOME when I don't have a microscope, and tomorrow's home town victory doesn't cause VICTORY! when I'm not standing near a wormhole).

In the rest of this section, I just want to briefly outline a more general strategy of defusing the less than savory implications of my view. The essence of the strategy is to deny that the semantic value of an expression is determinative of its cognitive significance. An expression's being meaningful is no reason to suppose that it has a meaning, as loopy as that sounds.

Recall first how I solved the mentalistic version of Frege's puzzle: I pled syntax. The reason you can believe that Hesperus is bright without believing that Phosphorus is bright is that you can't derive the one from the other without additional premises (like that Hesperus is Phosphorus). They entail[46] one another, sure—but we don't have access to entailment relations except insofar as they're mediated by derivability relations. That's just the principle that the mind isn't magical.

[46] I'm using 'entail' in this context as follows: Γ entails φ iff whenever the members of Γ are true in some world w, φ is true in w.

Similarly, I'm inclined to think that one can't tell by looking at a sentence whether it has content. You can look at it and tell what would have to be true for it to be true (that is, what other sentences you could deduce it from) and what would be true were it true (what other sentences you can deduce from it). But you can't tell whether it can be true in the first place—whether it has a content at all. Once you accept the explanation that 'unicorns fly' can be contentless but meaningful because it doesn't wear its impossibility on its syntax, it's very difficult to see why you couldn't then go on to tell a similar story for '2 + 2 = 5'.

I fear I'm being opaque. Let me give it one more go. Someone who's in the business of determining the data structures and algorithms used by the mind can tell you which representations are cognitively distinct for an agent, because she knows how those representations are processed and in particular whether they'll be processed identically or not. To a lesser extent, the agents themselves are possessed of this information, in that they have, upon occasion, equivocal evidence concerning whether they process two representations identically. But neither these 'syntactic' cognitive psychologists nor the agents they investigate is in any position to determine what the contents of the representations under investigation are, nor indeed whether they have contents in the first place. That task requires a substantive metasemantic theory and an investigation of how the agents interact with the world, not merely an investigation of how they manipulate data structures. It thus follows that agents, even highly sophisticated human agents, are in no position to protest that their representations mean X and not Y, or X rather than nothing, when theory and evidence say otherwise. (I should put the case stronger. They

don't get *any* say in the matter.) By my own lights I've done myself a disservice in talking of 'biting bullets', as though the complainants had any to shoot my way.

14. Conclusion

The theme of this dissertation is semantic uniformity. If some expressions are non-descriptive, all of them are; if some are rigid, all of them are; and if some get their semantic value from the (appropriate) causal relations they bear to things, all of them do. It's not a popular position, but it's not unglamorous either. More importantly, *it may be true*.

Global Metasemantic Uniformity holds that all simple *and* complex expressions have their contents determined in the same way, and I endorsed a version of the claim that the way such contents were determined was via the causal/ informational/ or lawful relations those expressions bore to things. As bizarre as that claim may seem, I argued early on in the chapter (in the argument from causal isolation) that the causal/ informational/ lawful version of Local Metasemantic Uniformity, the claim that all *simple* expressions have their contents determined by the causal/ informational/ lawful relations they bear to things, entails its Global counterpart. Local Uniformity is harder to deny and, more to the point, it *isn't* denied by metasemanticists like Fodor[47] and Dretske.

The other principal considerations I discussed were, first, that Global Uniformity is Ontologically Neutral in a way that CMT isn't. For a complex expression to legitimately have a semantic value it is not necessary, on my view, that its constituent expressions have semantic values. Unfortunately, examples showing the utility of such

[47] I've heard rumors that Fodor has gone conceptual role on the connectives. Sic transit etc.

neutrality that can't be labeled idioms are difficult to construct. I mentioned measure phrases above. Perhaps another example is the productive use of locutions typically thought of as existentially committing when combined with things we're intuitively not committed to, like sakes (as in 'for the sake of illustration' with the definite article 'the'). So I can say things like: 'John provided the example for the sake of illustration' refers to the state of affairs in which John provided the example and did so because he thought it illustrated his point, whereas your standard CMT-endorser has a slew of hoops to jump through to sweep sakes those things that don't exist under the rug.

Another case (and this shades into the second consideration) is that of intentional transitives[48]—for example, 'John is hunting unicorns.' The CMT-endorser might feel obligated to first recover a clausal complement (Quine (1956), den Dikken, Larson & Ludlow (1996)) and then solve Frege's puzzle in some way that itself respects CMT (no small task in itself). I, on the other hand, don't need to say anything of the sort: I can have 'John is hunting unicorns' express a relation between John and a proposition without first recovering a clausal complement and I don't need to solve Frege's puzzle, because it's a puzzle exclusively reserved for CMT-endorsers (and that was the second consideration, in case you missed it).

The final consideration was just how right all this seems. CMT-endorsers hate their own theory so bad they throw it under the bus at the slightest provocation. The theory tells them what to do with 'John is hunting unicorns': it's got meaningless parts, so it's meaningless. It tells them what to do with minimal Frege case pairs: parts with the same meanings, so wholes with the same meanings. It tells them what to do with

[48] As a sad historical fact, these are often called 'intensional transitives' as though they created intensional contexts and not hyperintensional ones.

sentences that existentially quantify over sakes: false, all of them. But compositionalists are naturally self-loathers I suppose. In each of these cases, they do what *my* theory tells them to do: look for the state of affairs that appropriately causes tokenings of the expression in question and take that to be the semantic value. They just persist in being perverse about it and monkey with the syntax and hyperintensionalize the semantics until it looks like their theory got it right (post hoc I might add. From the crowd that's so fond of triumphal proclamations of its predictive successes, too.)

After making my case, I looked at some objections. First, I considered whether the putative computability of the contents of complex mentalese expressions was any reason to accept CMT. My answer was no: just because the semantic values of complex expressions are computable, it doesn't mean they're computable from their constitution properties; and just because they're computable from their constitution properties, it doesn't mean they're metaphysically determined by their constitution properties. Furthermore, the principal argument for supposing that the semantic values of complex natural language expressions are computable (namely, that we compute them) doesn't extend to the semantic values of complex mentalese expressions (because we typically don't compute them).

Second, I looked at arguments for CMT from learnability and understandability. The basic premise in such arguments is that our finite minds' capacity to understand an infinite number of complex mentalese expressions requires that the semantic content of those expressions be computable from their constitution properties. Again, even if this were so, it doesn't then follow that those contents would then metaphysically depend upon those constitution properties. But, setting that issue aside, the premise itself is false:

any requirement that understanding the language of thought always requires computing the semantic values of complex expressions from their constitution properties leads to an infinite regress.

Third, I took on systematicity. I considered two claims that were supposed to follow from the putative systematicity of mentalese: that complex mentalese expressions have constituent structure and that the contents of complex mentalese expressions literally contain the contents of their simple constituents. The first claim I endorsed, and pointed out that it in no way conflicted with my theory. The second claim I rejected, and pointed out that it was false (on the evidence of Johnson (2006)). I should add (this was in a footnote) that it's doubtful whether natural language and for that matter mentalese is systematic: see Johnson (2004). (By the way, I'm no relation of Kent Johnson, insofar as I'm aware.)

Finally, we looked at some problem cases involving complex expressions that seem to be contentful, yet cannot stand in the appropriate causal relations to their seeming contents. In none of these cases were the problems unique to me: simple causal metasemantic theories faced the same worries. I argued that the problem cases weren't so wide-reaching as they at first appeared, and I suggested that perhaps the data (the seeming contentfulness of the expressions) might not even be relevant for metasemantic theorizing, in that there exists no plausible mechanism by which seeming contentfulness indicates actual contentfulness.

At the same time I recognize that there's a lacuna (read: gaping hole) in what's been presented here. I haven't solved the problem of intentionality. I haven't shown causal metasemantic theories to be better than their rivals. I've argued that if we accept a

causal metasemantic theory for simple expressions, we should do likewise for complex expressions, and thus abandon CMT. I'm not amiss in pointing out which evidence I take to be irrelevant to the truth of that conditional (intuitions about the contentfulness of certain expressions). But I do recognize (genuinely) that the antecedent has not been established. If it sounded the right side of barbarous, I'd have probably called this paper "Against Compositionality (As a Metasemantic Thesis) (For Those Who Already Hold a Causal Metasemantic Account for Simple Expressions)."

That's not to say that there aren't still reasons to reject CMT even for non-causal theorists. Global Uniformity, Ontological Neutrality, and a fruitful approach to Frege cases are benefits that accrue to anyone willing to deny CMT, and it strikes me that conceptual role semanticists or their ilk, though no allies of mine, may yet find fodder for their own theories in the foregoing. Ultimately, what I'm trying to do here is put a position on the table, to defend the viability of rejecting CMT and similar positions it entails (CCT and SST). CMT may be right at the end of the day, but the arguments for it are unpersuasive, and the case against it holds water, as far as I can see.

REFERENCES

Burge, T. (1973). "Reference and proper names," *The Journal of Philosophy*, 70, 425-439.

Burge, T. (1979). "Individualism and the mental," *Midwest Studies in Philosophy*, 4, 73-121.

Carnap, R. (1988). *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

Cook, M. (1980). "If 'cat' is a rigid designator, what does it designate?" *Philosophical Studies*, 37, 61-64.

Cumming, S. (2007). *Proper Nouns*. (Doctoral Dissertation). Rutgers, The State University of New Jersey.

Davidson, D. (1984). "Theories of meaning and learnable languages." In *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

den Dikken, M., Larson, R., and Ludlow, P. (1996). "Intensional 'transitive' verbs and concealed complement clauses," *Rivista di Linguistica*, 8, 331-348

Devitt, M. and Sterelny, K. (1999). *Language and Reality*. 2nd Edition, Cambridge, MA: The MIT Press.

Devitt, M. (2005). "Rigid Application," *Philosophical Studies*, 125, 139-165.

Donnellan, K. (1983). "Kripke and Putnam on natural kind terms," in C. Ginet & S. Shoemaker (eds.), *Knowledge and Mind*. Oxford: Oxford University Press.

Donnellan, K. (1993). "There is a word for that kind of thing: an investigation of two thought experiments," *Philosophical Perspectives*, 7, 155-171.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge: The MIT Press.

Drestke, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge: The MIT Press.

Evans, G. (1985). "Reference and contingency." In *Collected Papers*. Oxford: Oxford University Press.

Fodor, J. (1990). "A theory of content II." In *A Theory of Content and Other Essays*. Cambridge: The MIT Press.

Fodor, J. (2004). "Water's water everywhere," *London Review of Books*, 26, 17-19.

Fodor, J. and Lepore, E. (2001). "Why compositionality won't go away: reflections on Horwich's 'deflationary' theory," *Ratio*, 14, 350-368.

Fodor, J. and Pylyshyn, Z. (1988). "Connectionism and cognitive architecture: a critical analysis," *Cognition*, 28, 3-71.

Frege, G. (2006). "On sense and Nominatum," in A. Martinich (ed.) *The Philosophy of Language, Fifth Edition*. Oxford: Oxford University Press.

Gómez-Torrente, M. (2006). "Rigidity and essentiality," *Mind*, 115, 227-259.

Groenendijk, J. & Stokhof, M. (1991). "Dynamic predicate logic," *Linguistics and Philosophy*, 14, 39-100.

Johnson, K. (2004). "On the systematicity of language and thought," *Journal of Philosophy*, 101, 111-139.

Johnson, K. (2006). "On the nature of reverse compositionality," *Erkenntnis*, 64, 37-60.

Kamp, H. (1981). "A theory of truth and semantic representation," in J. Groenendijk, T. Janssen & M. Stokhof (eds.), *Formal Methods in the Study of Language*. Amsterdam: Mathematical Centre Tracts 135.

Kaplan, D. (2004). "Demonstratives," in S. Davis & B. Gillon (eds.), *Semantics: A Reader*. Oxford: Oxford University Press.

Katz, J. (1975). "Logic and language: an examination of recent criticisms of intensionalism," in K. Gunderson (ed.), *Language, Mind, and Knowledge (Minnesota Studies in the Philosophy of Science Vol. 7*). Minneapolis: University of Minnesota Press.

Katz, J. (1979). "The neoclassical theory of reference," in P. French, T. Uehling, Jr., and H. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press.

King, J. (2001). *Complex Demonstratives: A Quantificational Account*. Cambridge: The MIT Press.

King, J. (2001b). "Remarks on the syntax and semantics of day designators," *Philosophical Perspectives*, 15, 291-333.

King, J. (2007). *The Nature and Structure of Content*. Oxford: Oxford University Press.

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

LaPorte, J. (1997). "Essential membership," *Philosophy of Science,* 64, 96-112.

LaPorte, J. (2000). "Rigidity and kind," *Philosophical Studies*, 97, 293-316.

Lewis, D. (1980). "Mad pain and Martian pain," in N. Block (ed.) *Readings in the Philosophy of Psychology, Vol. I*. Cambridge: Harvard University Press.

Lewis, D. (2001). *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

Linsky, B. (2006). "General terms as rigid designators," *Philosophical Studies*, 128, 655-667.

Locke, J. (1979). *An Essay Concerning Human Understanding*. P. Nidditch (ed.), Oxford: Oxford University Press.

López de Sa., D. (2008). "Rigidity for predicates and the trivialization problem," *Philosophers' Imprint*, 8, 1-13.

Martí, G. (2004). "Rigidity and general terms," *Proceedings of the Aristotelian Society*, 104, 131-148.

Martí, G. and Martínez-Fernández. (2010). "General terms as designators: a defense of the view," in H. Beebee & N. Sabbarton-Leary (eds.), *The Semantics and Metaphysics of Natural Kinds*. New York: Routledge.

McGinn, C. (1982). "Rigid designation and semantic value," *The Philosophical Quarterly*, 32, 97-115.

Millikan, R. (1989). "Biosemantics," *Journal of Philosophy*, 86, 281-97.

Neale, S. (1990). *Descriptions*. Cambridge: The MIT Press.

Putnam, H. (1973). "Meaning and reference," *The Journal of Philosophy*, 70, 699-711.

Putnam, H. (1975). "The logic of quantum mechanics." In *Mathematics, Matter and Method*. Cambridge: The MIT Press.

Putnam, H. (1975b). "The meaning of 'meaning'." In *Mind, Language, and Reality*, Cambridge: Cambridge University Press.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge: The MIT Press.

Quine, W. (1956). "Quantifiers and propositional attitudes," *The Philosophical Review*, 53, 177-187.

Quine, W. (1969). "Natural kinds" in *Ontological Relativity & Other Essays*. New York: Columbia University Press.

Qvarnström, B. (1988). "The meaning of natural kind words," *Metalogicon*, 1, 58-70.

Salmon, N. (2005). *Reference and Essence, 2nd Edition*. Amherst: Prometheus Books.

Salmon, N. (2005b). "Are general terms rigid?" *Linguistics and Philosophy*, 28, 117-134.

Schwartz, S. (1980). "Natural kinds and nominal kinds," *Mind*, 89, 182-195.

Schwartz, S. (2002). "Kinds, general terms, and rigidity: a reply to LaPorte," *Philosophical Studies*, 109, 265-277.

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

Soames, S. (2004). "Reply to Ezcurdia and Gómez-Torrente," *Crítica, Revista Hispanoamericana de Filosofía*, 36, 83–114

Stalnaker, R. (1987). *Inquiry*. Cambridge, The MIT Press.

Stanley, J. and Williamson, T. (2001). "Knowing how," *The Journal of Philosophy*, 98, 411-44.

Sullivan, A. (2007). "Rigid designation and semantic structure," *Philosophers' Imprint*, 7, 1-22.

Szabó, Z. (2000). "Compositionality as supervenience," *Linguistics and Philosophy*, 23, 475-505.

Wiggins, D. (1980). "Putnam's doctrine of natural kind words and Frege's doctrines of sense, reference, and extension: can they cohere?" in P. Clark and B. Hale (eds.), *Reading Putnam*. Oxford: Basil Blackwell.