12

Semantics as Model-Based Science

Seth Yalcin

What is the right way of thinking about the metalanguage in which natural language semantics takes place?

I will eventually get to the idea that the best way of understanding the role of this language involves seeing natural language semantics as a model-based science (using 'model' in the general sense of a scientific model, not in the technical sense of model theory). To get there, I will first critique some standard ways of thinking about how the metalanguage of semantics is to be understood.

12.1 Heim and Kratzer on Showing Meanings

At the beginning of their excellent and influential textbook in natural language semantics, Heim and Kratzer (1998) are concerned to introduce the reader to a basic extensional semantics for a small fragment of English. Their overall approach overlaps in important respects with the standard model-theoretic semantics for first-order logic, though they work with more realistic syntactic structures and don't state their semantics syncategorematically. A domain of individuals is assumed. Proper names are treated like constant terms: their semantic values are entities from the domain. The semantic values of monadic predicates are taken to be functions from individuals to truth values. Using boldface as a device of quotation, Heim and Kratzer give lexical entries like this:

[[**Ann**]] = Ann
[[**smokes**]] = $\lambda x.x$ smokes

With the help of explicit rules of semantic composition (like a basic rule of functional application), their system extends semantic values to complex expressions. For example, early on, Heim and Kratzer consider simple derivations along these lines:

[[**Ann smokes**]] = [[**smokes**]]([[**Ann**]])
= $\lambda x.x$ smokes (Ann)
= 1 iff Ann smokes

Discussing the import of such derivations, Heim and Kratzer confront a question that arises owing to the extensional character of their semantics. They observe that their semantic value for 'smokes' is basically equivalent to the set of (actual) smokers. Suppose the actual smokers consist of just Ann and Jan. Then, given Heim and Kratzer's approach, we could have written the semantic value of 'smokes' equivalently as:

$[\![\textbf{smokes}]\!] = \lambda x.x \in \{\text{Ann, Jan}\}$

And then we might have calculated as follows:

$$\begin{aligned}[\![\textbf{Ann smokes}]\!] &= [\![\textbf{smokes}]\!]([\![\textbf{Ann}]\!]) \\ &= \lambda x.x \in \{\text{Ann, Jan}\}\ (\text{Ann}) \\ &= 1 \text{ iff Ann} \in \{\text{Ann, Jan}\} \\ &= 1\end{aligned}$$

And this looks strange at best. The conception of predicative and sentential meanings on display here seems most implausible—especially, the idea that the semantic values of sentences are merely truth values. What to say to the beginner?

Heim and Kratzer confront the question directly. They first quote the following passage from Dummett (1973):

> It has become a standard complaint that Frege talks a great deal about the senses of expressions, but nowhere gives an account of what constitutes such a sense. This complaint is partly unfair: for Frege the sense of an expres-sion is the manner in which we determine its reference, and he tells us a great deal about the kind of reference possessed by expressions of different types, thereby specifying the form that the senses of such expressions must take . . . The sense of an expression is the mode of presentation of the referent: in saying what the referent is, we have to choose a particular way of saying this, a particular means of determining something as a referent. (227)

Building on Dummett, Heim and Kratzer continue:

> What Dummett says in this passage is that when specifying the extension (reference, *Bedeutung*) of an expression, we have to choose a particular way of presenting it, and it is this manner of presentation that might be considered the meaning (sense, *Sinn*) of the expression. The function that is the extension of a predicate can be presented by providing a condition or by displaying it in a table, for example. Only if we provide a condition do we choose a mode of presentation that 'shows' the meaning of the predicates and the sentences they occur in. Different ways of defining the same extensions, then, can make a theoretical difference. Not all choices yield a theory that pairs sentences with their truth-conditions. Hence not all choices lead to a theory of meaning. (1998: 22)

The crux of Heim and Kratzer's response is evidently to distinguish the meaning of an expression from its semantic value. The interpretation function, it turns out, does not after all map expressions to their meanings. The semantic value of an intransitive predicate is a certain function from individuals to truth values, but this isn't actually

the meaning of the predicate. Instead the meaning is something 'revealed' by stating the semantic value of the predicate in a quite particular kind of way. Similarly, the semantic value of a particular sentence is a truth value, but this truth value isn't the meaning of the sentence. The meaning of the sentence is something 'shown' by specifying its semantic value (a truth value) in a certain way. On this conception of things, one can take a correct semantic theory and render it incorrect without materially changing a single semantic value or composition rule.

I object. This is a step into darkness. Such a view is unstable and implausible, and should be rejected.

It is unstable because it doesn't apply in any clear and principled way to a great deal of meaningful language—including some of the language that Heim and Kratzer are foremost concerned with. Take quantification as an example. Later in the textbook, Heim and Kratzer introduce some celebrated ideas about how to give the semantics of quantificational determiners such as 'every'. They state a semantic value for 'every' along the following lines:

$[\![\textbf{every}]\!] = \lambda f_{et}.\lambda g_{et}.$for all x such that $f(x) = 1, g(x) = 1$

But it is notable that Heim and Kratzer do not pause to consider what might seem by the lights of their earlier methodological comments to be a pressing question: why be confident that this semantic value adequately 'shows' the meaning of the quantificational determiner? For obviously there are numerous extensionally equivalent alternatives—numerous other modes of presentation of this semantic value. For example:

$[\![\textbf{every}]\!] = \lambda f_{et}.\lambda g_{et}.$every x such that $f(x) = 1$ is such that $g(x) = 1$
$[\![\textbf{every}]\!] = \lambda f_{et}.\lambda g_{et}.$for all x, if $f(x) = 1$ then $g(x) = 1$
$[\![\textbf{every}]\!] = \lambda f_{et}.\lambda g_{et}.$for every x, $f(x) = 0$ or $g(x) = 1$
$[\![\textbf{every}]\!] = \lambda f_{et}.\lambda g_{et}.$for all x, $g(x) = 1$ or $f(x) = 0$

Should we take seriously a conception of semantics which allows that one theory might be superior to another extensionally equivalent theory, owing merely to the fact that the first theory but not the second was stated with a particular one of these entries rather than another? (Or is it that these all happen to be equally meaning-revelatory for 'every'? But by what leap of insight is that supposed to be evident?) I suspect that any semanticist would rightly regard it as a confused question which of these lexical entries 'shows' the meaning of 'every' best.

Why is it a confused question? Differing hypotheses about what the semantic value of a quantificational determiner is make predictions that can be evaluated—for example, predictions about relations of consistency and entailment between sentences involving that determiner, about truth-conditions, about judgments of felicity, and so on. Changing up only the way that a given semantic value for 'every' is 'shown' does not affect any of these these predictions—it does not affect what is, in ordinary semantic practice, thought to be the empirical content of that proposed semantic value.

Semantic theories that hypothesize exactly the same semantic values and composition rules and which vary only in how the semantic values are stated would seem to differ only hyperintensionally. Is there any precedent in science for favoring one theory over another one entirely owing to such hyperintensional differences? It may be correct in some cases to distinguish scientific theories which are intensionally equivalent. Two theories might agree in their predictions modally and yet differ in respect of relations of explanatory order, or of 'ground', or of metaphysical dependence. But it is clear that this is not what we are seeing suggested in Heim and Kratzer's remarks. What they are suggesting is the view that a semantic theory must, besides making correct predictions and having the usual theoretical virtues of simplicity and parsimony, be stated in a way that satisfies a special kind of insight—an insight into ways of specifying semantic values that are in some sense meaning-revelatory. But it hard to see why natural language semantics should involve this mysterious additional requirement. This seems to involve a kind of 'methodological dualism', the idea that the empirical study of meaning is subject to special restrictions or requirements not standardly recognized in ordinary scientific inquiry.

It might be replied that I am not fully appreciating the empirical content of a semantic theory on Heim and Kratzer's view (or some charitable reinterpretation of it). Perhaps the way to take their proposal is that a semantic theory, besides making the usual predications about truth-conditions, entailment, and the like, also makes a class of explicit predictions about which semantic clauses certain native speakers of a language will judge to be meaning-revelatory in the relevant sense. So their proposal is that among the predictions generated by a semantic theory are predictions of the form:

> English speakers of a certain dialect with suitable technical training will view
>
> [[**smokes**]] = $\lambda x.x$ smokes
>
> as meaning-revelatory.

Such a theory then differs in empirical content from another theory which otherwise assumes the same semantic values, e.g., a theory which entails:

> English speakers of a certain dialect with suitable technical training will view
>
> [[**smokes**]] = $\lambda x.x \in$ {Ann, Jan}
>
> as meaning-revelatory.

We can agree we now have a point of empirical divergence. But the implausibility of this general path is manifest. The idea of *being meaning-revelatory to a person with suitable technical training* is not a clear notion, as we have noted. But even if it were clear, why should semantic theory be understood as partly a theory of how speakers with a certain technical training are apt to respond to other clauses in that very theory? The weirdness of this proposal should be acknowledged. Natural language semantics is supposed to be (part of) a theory of what competent speakers know when they know

the meanings of the expressions of their language. It is not also somehow a theory of how linguists are apt to respond to certain technical clauses of natural language semantics itself.

It seems to me that any appeal this approach has rests on a certain confusion, namely the confusion that the word 'smokes' in the metalanguage is the very word 'smokes' of the object language. Assuming this, one might think native speaker judgments about the metalanguage are somehow probative of something in the explananda of a semantic theory for the object language. But this is a serious mistake about how to understand the metalanguage, as I will try to make clear.

An additional concern about primitive appeals to meaning-revelatoriness is that these appeals seem to presuppose just the distinctions we want a semantic theory to further explain and model. It is hard to see how a theory making brute appeal to facts of meaning-revelatoriness can ever be in position to provide a deeper explanation of those facts—of the sort we might achieve instead, for instance, by grounding these differences in differences of semantic value. Facts of semantic performance can be brought in as data to constrain theorizing about semantic competence. But the theorist does not somehow deploy her semantic competence directly in order to settle theoretical choices (choices framed and faced by the theorist, stated in theoretical terms)—much less choices between semantic theories that are completely equivalent except in mode of presentation. One settles theoretical choices by weighing their relative simplicity, parsimony, explanatory power, descriptive coverage, etc.—by applying ordinary norms of scientific theorizing. Theoretical choices are not settled by unexplained events of meaning-revelation. One does not as it were present a technical theory before the mind, and confirm or disconfirm it merely letting the mind directly reveal its truth or falsity, like an oracle. But it is not easy to see how Heim and Kratzer's methodology of theory selection significantly differs from this, once we get to theories differing only in how semantic values are stated.

The natural alternative methodological course is not to recognize any distinction between the meaning of an expression and its semantic value. If a proposal about the semantic value of an expression leads to implausible results, that is just a sign that the proposal is wrong, or that other assumptions in the theory need adjustment. It is not a sign that the proposal needs to be swapped with an intensionally equivalent but hyperintensionally distinct proposal. The compositional semantic value of an expression just is our theoretical analysis of the meaning of the expression—subject to revision by further inquiry, of course, but that goes without saying. In any science, we see the replacement of everyday notions with theoretical counterpart notions; so it is with the move from 'meaning' to 'semantic value'.

The question then arises why Heim and Kratzer seem to suppose that a different attitude is supposed to be appropriate to towards the statement of the semantic value of predicates, or of sentences. The crucial aspect of the quote above seems to be this:

the extension of a predicate can be presented by providing a condition or by displaying it in a table, for example. Only if we provide a condition do we choose a mode of presentation that 'shows' the meaning of the predicates and the sentences they occur in. (22)

Why do Heim and Kratzer say this? One reason may be the following. If the meaning of an expression were identified with its semantic value, then Heim and Kratzer's opening assumptions about the meanings of sentences and of predicates would just manifestly be nonstarters. It does not take much argument to establish that extensionally equivalent predicates may have distinct meanings, or that there are more than two possible meanings for sentences. So meaning must reside somewhere beyond semantic value, if Heim and Kratzer's opening extensional semantic theory is not to be a nonstarter.

But it seems a better reaction would be to acknowledge the severe limitations of a purely extensional theory at the outset—to grant that in fundamental respects, such a theory really is a nonstarter. The conclusion to draw from the problem they raise is not that meaning must reside somewhere beyond semantic value; it is that the semantic values initially postulated are not fruitful, because too coarse-grained. We need richer semantic values to capture the sorts of distinctions that need distinguishing. To respond to the problem by appealing instead to hyperintensional distinctions between equivalent semantic theories seriously obscures the fundamental failures of the semantic values posited by the extensional theory. The appearance that the initial problem about the meaning of predicates and of sentences goes away once we restate the semantic value in a certain way is an illusion, one encouraged by Heim and Kratzer's casual use of the object language in the metalanguage. That leads the incautious reader to read more into the metalanguage than is really there—and in particular it makes it easy to tacitly appeal to the very aspects of our semantic competence that were supposed to be explicated by the theory. (More on this shortly.)

A better response to the problem would simply be to introduce intensional resources (possible worlds or situations) at the start, as a beginning at fixing the problem, at delivering a semantic theory that can make at least a minimal range of the distinctions between semantic values that need to be distinguished. It is sometimes thought that intensional resources should only be introduced in semantic theory when one comes to analyzing intensional language, like modal operators. But even if our language were entirely free of intensional devices, a semantic theory with intensional resources would still plausibly be necessary to characterize the contribution of meaning to the transfer of information in communication, and in order to adequately capture entailment relations between sentences. At least, this is so if some version of a broadly truth-conditional approach to meaning is assumed (the kind of approach I restrict attention to in this chapter). The notion of the truth-condition of a sentence is a fundamentally modal notion: it concerns what *would* be

were the sentence true.[1] A 'condition' denuded of modality is not a condition. If we theorize in a way that accords a central place to truth-conditions, it is hard to see how one can escape the use, in one's theory, of things like possible worlds or situations (or the use of some allied modal ideology).

The first sentence of Heim and Kratzer (1998) is a familiar mantra: 'To know the meaning of a sentence is to know its truth-conditions'. Since such knowledge is modal in character, plausibly modality should be part of the architecture of the theory of this knowledge from the outset.

12.2 Disquotation

When Heim and Kratzer, in their extensional semantics, end a semantic derivation in this kind of way:

[[**Ann smokes**]] = 1 iff Ann smokes

an illusion is created that the object language sentence has been associated with truth-conditions—with the way things would have to be, were it true that Ann smokes. The illusion is encouraged partly by a casual use of the object language in the metalanguage (and also, perhaps, by the ease with which the material biconditional can be misheard has having modal import). We interpret the metalanguage sentence 'Ann smokes' in the ordinary, modally rich way that it is understood in English. But this involves an illicit use of a dimension of our competence, a dimension that the theory itself was supposed to model and explain. The extensional semantic theory Heim and Kratzer offer does not associate 'Ann smokes' with the condition on possible states of the world which obtains just in case Ann smokes. Truth-conditions adequate to the slogan 'To know the meaning of a sentence is to know its truth-conditions' are not secured here.

In Heim and Kratzer (1998) there is a striking absence of explicit reference to models, the fundamental notion in the semantics of formal languages. One might think this was a pedagogical decision—a decision to keep the formalism to the minimum for the beginner. But there is evidently a deeper motivation at work. On their first page, they suggest that a theory of meaning produces statements of the following form:

The sentence '_____' is true iff _____.

[1] Other textbooks in semantics make this elementary point rather explicitly; see, e.g., Dowty et al. (1980: 12); Jacobson (2014: 32). This point is of course at the heart of the so-called 'Foster Problem' for Davidsonian approaches in semantics; see Foster (1976: 11).

(It is worth flagging here that contemporary work recognizes a conceptual distinction between compositional semantic value and content. Each of these notions might have affiliate models of truth-conditions. See Yalcin 2014 for further discussion.)

where 'iff' is the material biconditional, and where the gaps might in principle be replaced with the very same sentence of English ('There is a bag of potatoes in my pantry', to give their example). Obviously this is the T-schema, tracing back to Tarski (1936, 1956). They next approvingly cite a famous passage from Davidson (1967a):

> The theory reveals nothing new about the conditions under which an individual sentence is true; it does not make those conditions any clearer than the sentence itself does. The work of the theory is in relating the known truth conditions of each sentence to those aspects ('words') of the sentence that recur in other sentences, and can be assigned identical roles in other sentences. Empirical power in such a theory depends on success in recovering the structure of a very complicated ability—the ability to speak and understand a language. (311)

Davidson's picture clearly exerts a nontrivial influence on Heim and Kratzer's metatheoretical reflections—he is the barely-concealed wizard behind the curtain.[2] But there are reasons to be seriously unhappy with what Davidson says here. Without pretending to be able to fully engage Davidson's intricate *Weltanschauung*, I will just highlight some places where I find difficulties.

First, if we adopted the suggestion that a (good) semantic theory for natural language 'reveals nothing new about the conditions under which an individual sentence is true', we would be compelled to abandon many of the most basic advances in natural language semantics. It is a familiar point that all sorts of 'new things' characteristically appear in the metalanguage used to interpret a sentence. For starters, typically we find reference to models, sets, domains, functions, types, and variable assignments. The detailed study of various particular language fragments often uncovers the need for further resources. With modals and conditionals, we arguably find a need for reference to possible worlds or situations, together with accessibility relations, orderings over worlds, selection functions, and the like. Analogous structure is typically postulated for tenses. With gradable adjectives, we arguably find a need to make reference a 'scale structure'—a set of points totally ordered along some 'dimension' (see e.g. Bartsch and Vennemann 1972; Kennedy 2007). With plural terms, we perhaps need to be able to talk about certain semilattices of individual sums of members of extension of predicates (Link 1983). There is of course the famous idea of interpreting most predicates as predicates of events (Davidson 1967b). The list of classic work in formal semantics in this vein could continue for quite a while. The shape of the best truth-condition for a given natural language sentence is generally *news*—news formulated in theoretical terms which are not part of natural language.

It is not easy to square this fact with the idea that disquotational truth-clauses constitute a central dimension of a good semantic theory, even for restricted fragments.

[2] It is not uncommon to distinguish Davidsonian approaches to truth-conditional semantics, which are taken to place a special emphasis on disquotation, from model-theoretic or Montogovian approaches to truth-conditional semantics, which place a special emphasis on truth relative to a model. Glanzberg (2014) locates Heim and Kratzer (1998) in the second tradition. But for the reasons reviewed here, I think it is less than clear whether this classification is accurate.

Theorists trying to square these sometimes appeal to the idea that any seemingly new elements mentioned in the truth-condition of a sentence must somehow correspond fairly directly to elements of the underlying syntactic structure being interpreted—so that in a sense, all news is syntactic news, and some abstract notion of disquotation is preserved. But it is difficult to make out what could recommend this methodological posture. An alternative idea, arguably implicit in most semantic theorizing, is the view that sentences like this:

(1) 'Grass is green' is true iff grass is green

in fact do not actually appear in any explanatory formal semantics for natural language. Instead what usually happens is that 'Grass is green', or a suitable structural description thereof, is associated with a set of points of evaluation within a specified model.[3]

Davidson and his followers would resist. Davidsonians hold that disquotational truth clauses do form a core component of (good, explanatory) natural language semantic theories. They might take the view just expressed to be glib about matters foundational. Maybe it will be suggested that the set of points I mention is a mere technical object whose role, when clearly apprehended, is simply to formalize, or somehow eventually get at, the ordinary-language condition *that grass is green*, a condition only 'graspable' by deploying one's natural language competence. Davidson seemed to think that natural language was in some deep sense inescapable even in theoretical explanations.[4] He also seemed to think that strict adherence to a theory producing disquotational T-sentences was the only clear path to a 'respectably empirical' semantic theory (Davidson 1973). Despite his regular invocations of Tarski, Davidson had an enduring suspicion of exactly the technical semantic notions (denotation, interpretation, satisfaction, truth at a model, etc.) Tarski had explicitly set out to render scientifically respectable. Indeed it is difficult to exaggerate the extent to which Davidson departs from Tarski in respect of his attitude towards semantic concepts. Tarski's relatively brief, informal discussions of the T-schema (in Tarski 1936, 1944) are heavily foregrounded by Davidson, while the actual recursive definitions of truth Tarski gives—with their essential use of technical semantic notions on the righthand

[3] I am harmlessly glossing over model-theoretic approaches that take a dynamic shape (e.g., Kamp 1981; Heim 1982; and much subsequent work).

[4] We find in Davidson (1973) the following remarks: 'interpreted formal systems are best seen as extensions or fragments of the natural language from which they borrow life' (71); 'if we understand our metalanguage, we are using a system of concepts and a language which is the one for which we *really* want a theory, for it is this richer system that is our natural one' (73); and 'The inevitable goal of semantic theory is a theory of a natural language couched in a natural language' (71). These claims are made as though defense were not necessary. I imagine that theorists like Frege, Tarski, and Chomsky, all of whom emphasize in various ways the limitations of natural language for scientific inquiry, would find these claims less evident. (Tarski in particular emphasizes that semantic clauses 'should not be regarded as definitions of the terms involved' (Tarski 1944: 357), and denies that these clauses involve the kind of conceptual circularity Davidson's disquotational picture entails.)

side of the biconditionals—are minimized, despite the fact that the latter is the obvious locus of Tarski's core achievements. Davidson sought a recursive theory of 'absolute' (unrelativized) truth, one eschewing the notion of truth at a model and vindicating the T-schema in unalloyed form (or in a form alloyed just enough to make room for context-sensitivity). But nowhere did he seriously attempt to show how such a theory could actually be given for something as elementary as quantification, the semantics of which Tarski famously compositionally modeled via a technical notion of satisfaction. Davidson acknowledged in various places the theoretical need to deploy technical semantic concepts—again, Tarski's work is unintelligible without them—but he continually downplayed their role. Insofar as notions like denotation, satisfaction, etc., seem necessary in semantic theorizing, they 'must be treated as so much theoretical construction' (Davidson 1973); they are just 'posits' or 'theoretical constructs' which are not 'open to direct confrontation with the evidence', and are to be understood in some instrumental, not fully serious way; in some philosophically important sense, 'we don't need' semantic concepts like reference, satisfaction, etc.—theories straightforwardly using these notions are in fact 'hopeless' (Davidson 1977). The relevant notion of 'direct confrontation with the evidence', apparently the litmus test of the real, was rooted in a basically Quinean behaviorist ideology according to which 'meaning is entirely determined by observable behavior, even readily observable behavior' (Davidson 1990)—an ideology invoking evidential norms foreign to ordinary scientific inquiry, a point familiar in essentials since Chomsky 1959.[5] It is not surprising that Davidson nowhere produces a concrete compositional semantic theory for any nontrivial fragment of a natural language which vindicates a disquotational T-schema *modulo* context-sensitivity. The clearest contemporary notions of truth at a context are all stated in the context of a model structure (Kaplan 1977/1989 being the lead example); and further, it now widely recognized that even this technical notion is not plausibly compositional in the relevant sense (Lewis 1980; Rabern 2013; see also Ninan 2010; Rabern 2012; Yli-Vakkuri 2013; Yalcin 2014). One can state a

[5] Related, Davidson seemed to hold the view that technical semantic concepts (semantic value, denotation, satisfaction, etc.) could possess full scientific legitimacy in application to natural language only if it we could say how facts of semantic value (etc.) were supposed to be reducible to facts characterized in nonsemantic terms (see, e.g., Davidson 1977). For criticism of this foundations-first kind of methodology, see Yalcin (2014).

A certain conflation of epistemology and metaphysics also seems to encourage his shrugging-off of semantic concepts as serious: from the fact that empirical confirmation of a semantic theory might come substantially by way of judgments about truth-conditions, brought to bear on the theory holistically—an epistemological fact—Davidson (in, e.g., Davidson 1977) seems to infer, incorrectly, that he is free to reject the view that the truth-condition of a sentence is metaphysically grounded in facts about the semantic values of the parts of the sentence. Davidson seemed in some places to reach for the view that natural language semantic reality is given, not by a single Tarskian semantic theory, but rather by a wide space of such semantic theories related by some kind of symmetry transformation wherein semantic values are radically permuted. I am unaware of arguments for this position in Davidson not ultimately rooted in behaviorist ideology. *Prima facie*, it is unclear how a theory incorporating such radical indeterminacy could possibly cohere with the idea of taking the compositionality of meaning seriously as one component of the productivity property of natural language.

semantic theory without making the underlying model structure perpetually explicit (cf. Heim and Kratzer 1998), but obviously, not mentioning it doesn't make it go away. Leaving this relativization implicit can produce the illusion that an unrelativized notion of truth has been recursively specified when in fact no such thing has been achieved.

Moving on, maybe some strand of neo-Davidsonian will concede that semantic theory does after all deliver some kind of news on the righthand side when it comes the 'closed class' expressions that form the 'functional' core of the language. Still, it might be maintained, when it comes to giving the semantics of 'open class' expressions like predicates and common nouns, our metalanguage will have to use the very words being interpreted. (Or translations of those words in another natural language.) This is illustrated by Heim and Kratzer's way of stating their semantic value for 'smokes' (repeated):

[[**smokes**]] = $\lambda x.x$ smokes

We see something similar in another influential (and explicitly Davidsonian) semantics textbook, Larson and Segal (1995). This textbook assigns semantic values to predicates like this:

Val(x, **ponders**) iff x ponders

In both cases, the expression being interpreted is used in the metalanguage to articulate its own interpretation. Semantic competence with the expression being interpreted is assumed of the reader of the metalanguage.

These features are objectionable, however. One should sense a disturbance in the force when a theory which is supposed to explain what it is to be semantically competent with an expression actually requires the theorist possess this competence herself, and further requires the theorist to deploy that competence in grasping the theoretical 'explanation' delivered. Insofar as semantically competence with the expression is presupposed to some extent in the metalanguage and is expected to do theoretical work, explanatory power is to that extent compromised. This is just a case of begging the question.[6, 7]

Fortunately, the value of the theoretical work performed by semantic values like those described above owes to the extent to which the righthand sides can be understood as *not* depending on competence with ordinary language or ordinary concepts. And the righthand sides displayed above are really considerably more remote in their

[6] Many philosophers of language undergo years of radical interpretation therapy in order to unsee this problem.

[7] Glanzberg (2014) raises a similar worry: he suggests that to the extent that semantic theories rely on disquotation, 'they threaten to become explanatorily vacuous' (269). Glanzberg does not frame is point as one about disquotational theories begging the question, but this framing seems compatible with his discussion.

content from anything expressible in ordinary language than Heim and Kratzer, or Larson and Segal, make it appear.

12.3 The Righthand Side Is Not a Translation of the Lefthand Side

Let me elaborate. As semantic theorists, we try to get at the meaning of (e.g.) 'smokes' by constructing a theoretical model of its semantic features.[8] Successful theorizing virtually always requires moving beyond ordinary thought and talk, and calls for recourse to technical notions and distinctions; we expect our ultimate theory to have many parts that are only loosely and inexactly describable in ordinary terms (as Frege and Tarski continually emphasized). One property that a predicate has in virtue of, or as a component of, its meaning, we postulate, is an extension. Inspired by the standard semantics of first-order logic, we model an extension as a subset of a set of individuals—a domain—the latter given in the context of some model; or we can model it as a function which maps individuals from the domain into truth values. From a theoretical point of view, the individuals in the domain only have whatever structure we find a need to postulate in order to capture differences of meaning. The domain of individuals does not literally contain objects from the manifest image of common-sense, things like Saul Kripke and London—rather, it contains abstract points whose theoretical role is to model aspects of the meaning of 'Saul Kripke' and 'London' (on some approaches to names, at least). The idea of a set of such individuals—basically, a set of abstract points—is in principle intelligible to the theorist in complete abstraction from any competence with the object language predicate whose meaning it is being used to model. We use the set-theoretical features of this extension to model logico-semantic relations between this predicate and others. A set of individuals in this sense does not metaphysically depend on any facts about meaning ('metaphysical dependence' in roughly the sense of Rosen 2010). The independent intelligibility and metaphysical independence of this object is important. We are modeling a dimension of the meaning of a predicate with a feature that is not itself essentially semantic in nature, one we can in principle understand without presupposing the very competence we are trying to give a theory of.[9]

Now as already noted, the semantic features of 'smokes' go far beyond its extension. To that extent, the above lexical entry for 'smokes' falls well short. The use of 'smokes' in the metalanguage can obscure this. We understand that word in the metalanguage

[8] Here I mean 'model' in the general sense of a scientific model. Of course, in natural language semantics, the models in question are typically built with tools from the mathematical theory of models, which uses 'model' in a separate and technical sense. More on this below.

[9] Of course, substantive theorizing in virtually any science often involves postulating models with arbitrary features; in such cases, we will often want to say that the underlying reality being modeled is really given by some kind of symmetry transformation on the model. My point here does not require commitment to the idea that a particular set-theoretic object is *the* semantic value of a given expression.

in all its ordinary semantic richness, deploying our semantic competence with the word; but none of this richness is actually captured by the semantic value stated above save its extension. Thus if we are not careful, we are apt to see much more on the righthand side than is really there.

I am saying that Heim and Kratzer are mistaken in assuming that the occurrence of 'smokes' on the righthand side of the lexical entry above is the very word 'smokes' being interpreted—a point Tarski seems to make in a related context (see fn. 4). Another way into this point is to observe that English words like 'smokes' simply do not occur grammatically in lambda abstracts. (Nor do they form open sentences with the variables of a formal language, as Larson and Segal (1995) assume.) We only understand 'smokes' on the right hand side insofar as we abstract away from core features of the ordinary word 'smokes'. We can (and do) do that, obviously, but it is important to recognize the abstraction involved. This is not hairsplitting. What I am emphasizing here is the degree to which a kind of theoretical understanding is in play on the righthand side even for the words that superficially seem to belong to the object language.[10]

The point here is especially evident in Heim and Kratzer's last chapter, when intensional resources finally appear on the stage. In this chapter, Heim and Kratzer upgrade predicative semantic values with possible worlds. The semantic value of 'smokes' is given like this:

$$[\![\textbf{smokes}]\!]^w = \lambda x.x \text{ smokes in } w$$

Is the word 'smokes' as it appears on the righthand side here the familiar 'smokes' of English, the word occurring within the interpretation brackets? It clearly is not. Rather, it is a piece of technical vocabulary, one which grammatically requires combination with a variable that ranges over individuals and a variable that ranges over possible worlds. The ordinary word 'smokes' plausibly does not have anything like this grammar.[11] The main point, however, is that *whether or not the English word has this grammar is an open empirical question, whereas there is no open empirical question how the word 'smokes' in the metalanguage as deployed here works*. It works the way we as theorists stipulate. It is used to name a certain possible worlds intension.

Let me repeat this point. There is a kind of open question argument that can be used to see that the object language occurrence of 'smokes' above is not synonymous with its metalanguage counterpart. There is a substantive empirical question what the semantic value of the English word 'smokes' is. But there is no comparable substantive empirical question question about what the homonymous 'smokes' of the metalanguage above means, or how it works. So these homonyms are not synonyms.

[10] In this section of the chapter, I seem to differ with Glanzberg (2014), who holds that disquotation plays a larger (if ultimately nonexplanatory) role in semantic practice of the model-theoretic stripe than I do.

[11] For recent work on world/situation variables in natural language syntax, see Keshet (2008, 2010); Schwarz (2012).

The English predicate 'smokes' can apparently optionally combine with a locative phrase, as in 'Jane smokes in her apartment', but it would of course be a complete confusion to understand the use of 'in' in the metalanguage above in the same way. The 'in *w*' of the metalanguage is not a locative phrase in any remotely normal sense; and its appearance is not optional but required. This is all fundamentally technical vocabulary—made up language—whose true import derives from the way it explicitly maps into some feature of our semantic model—in this case, into certain intensions. Our understanding of these features of models is of a theoretical nature. It does not fundamentally require competence with the object language it is used to interpret.

The use of the word 'interpret' here can cause confusion, because this word, like 'translate', is also used to describe ordinary intuitive relations between words from distinct natural languages. On the picture of Davidson (1967a), a semantic theory associates sentences of the object language with a translation of it in the metalanguage. The sense of 'translation' Davidson intended is quite literal—it is the same sense in play as when we say that the German 'Katze' can be translated into English as 'cat'. But no such notion of translation is in play in understanding what is going on when semantic values are stated in semantic theory. The expression '$\lambda x.x$ smokes in w' is not a 'translation' of 'smokes'—certainly not in anything remotely like the sense that 'cat' translates 'Katze'. Nor is the object language 'smokes' equivalent in meaning to the metalanguage word 'smokes', for the reasons given in the last two paragraphs. The expression '$\lambda x.x$ smokes in w' picks out a formal object that (we hypothesize) serves to model the semantic contribution of 'smokes' given the appropriate theoretical context and some restricted empirical domain. Our understanding of '$\lambda x.x$ smokes in w' derives from a kind of theoretical understanding; it requires knowledge of semantic theory to understand. The English word 'smokes' does not.

12.4 Ordinary Notions Not the Material for Scientific Theorizing

Maybe I will seem to some to be stubbornly evading what would appear to be perfectly ordinary and natural characterizations of routine practice in natural language semantics. It could be granted that

$\lambda x.x$ smokes in w

picks out a possible worlds intension relative to a model—a certain kind of technical, theoretical object, in the sense that competent speakers needn't have any clue what this is; and it could be agreed that 'smokes' here is not exactly the 'smokes' of English. Still, one might think, what is fundamentally happening here is that we are picking out a function from worlds to the set of those who *smoke* in that world; and when we say 'smoke' here, we mean that in the ordinary English sense. Thus to articulate what this intension does, we use our competence with the expression interpreted.

This is indeed how intensions are sometimes characterized in semantics as a matter of informal practice. But it is important to recognize that insofar as the theorist helps herself in the metalanguage to talk of those who smoke, using this talk to articulate the meaning of 'smokes', something fundamental is going unexplained. It is not plausible to assume that the ordinary, common-sense notion of smoking, the one voiced by 'smokes', is among the resources that a theorist concerned with the meaning of 'smokes' can legitimately help herself to. Rather, that notion is ultimately part of the explananda. We cannot glibly assume that there is some mind- and language-independent joint in nature that separates those who smoke from the rest—that this distinction can be understood in complete abstraction from human language and cognition, the science of which natural language semantics is a chapter of. If there were such a joint, it might be legitimate for the theorist to appeal to it in her explanations of natural language meaning. But there is little reason to doubt that such ordinary notions are caked with features of parochial human thought and cognition—features that, on the face of it, are integral to their meaning, and hence features that it would be reasonable to expect semantic theory to play at least some role in illuminating.

Chomsky has made the basic point here in various places. Some passages from Chomsky (2000):

> Such [ordinary] notions as *desk* or *book* or *house*, let alone more 'abstract' ones, are not appropriate for naturalistic inquiry. Whether something is properly described as a desk, rather than a table or a hard bed, depends on its designer's intentions and the ways we and others (intend to) use it, among other factors. Books are concrete objects. We can refer to them as such ('the book weighs five pounds'), or from an abstract perspective ('who wrote the book?'; 'he wrote the book in his head, but then forgot about it'); or from both perspectives simultaneously ('the book he wrote weighed five pounds,' 'the book he is writing will weigh at least five pounds if it is ever published'). If I say 'that deck of cards, which is missing a Queen, is too worn to use,' that deck of cards is simultaneously taken to be a defective set and a strange sort of scattered 'concrete object,' surely not a mereological sum. The term *house* is used to refer to concrete objects, but from the standpoint of special human interests and goals and with curious properties. A house can be destroyed and rebuilt, like a city; London could be completely destroyed and rebuilt up the Thames in 1,000 years and still be London, under some circumstances. It is hard to imagine how these could be fit concepts for theoretical study of things, events, and processes in the natural world. (20–1)

> . . . not that there are no such things as desks, etc., but that in the domain where questions of realism arise in a serious way, in the context of the search for laws of nature, objects are not conceived from the peculiar perspectives provided by concepts of common-sense. (21)

> We can regard London with or without regard to its population: from one point of view, it is the same city if its people desert it; from another, we can say that London came to have a harsher feel to it through the Thatcher years, a comment on how people act and live. Referring to London, we can be talking about a location or area, people who sometimes live there, the air above it (but not too high), buildings, institutions, etc., in various combinations (as in *London is so unhappy, ugly, and polluted that it should be destroyed and rebuilt 100 miles away*, still being the same

city). Such terms as *London* are used to talk about the actual world, but there neither are nor are believed to be things-in-the-world with the properties of the intricate modes of reference that a city name encapsulates. (27)

Our evidently intricate and diverse modes of reference to aspects of reality (to London, books, houses, and so on) should be at least partly illuminated by our ultimate theory of meaning of the associated expressions (our theory of 'London', 'books', 'houses', etc.) and by our theories of affiliate aspects of cognition. Clarity is not achieved merely by declaring that the semantic value of 'London' is London (etc.). If explanations ended there, the sort of puzzling facts about reference Chomsky draws our attention to in these passages would need to be explained by a dubious metaphysics for cities, books, houses, and so on; and that would mislocate the problems. (The troubles here are not about London; they are about 'London'.) We may wish to table such problems if our theoretical interests in semantics lie elsewhere; that might be perfectly sensible. A simplistic assumption the effect that 'London' denotes a certain individual in a model may be adequately serviceable in many explanatory contexts. This isn't being denied. But we should not be under illusions to the effect that that these problems don't exist, or that they are not at least part of the explananda of semantic theory, or that they go away by saying that 'London' denotes London.

Part of what it is for a notion to be common-sensical is for it to strike us as natural and obvious. Being so close to view, it can be difficult to bring into focus what is problematic about using such notions in theoretical contexts. But it is a familiar point that successful scientific inquiry often requires one to adopt a theoretical perspective wherein the familiar and obvious comes to appear strange and problematic. Viewed through the appropriate theoretical lens, one interested in clarifying the nature of meaning in language and the associated underlying dimensions of cognition, ordinary commonplace notions will mostly strike us as puzzling creatures, dependent in complex ways on human interests, concerns, and modes of thinking in ways that cry out for explanation. I take it one of Chomsky's objectives with the examples above is to draw this out. We can try to build scientific theories (in psychology) *of* these notions—we can try to build theories articulating their role in our cognitive lives, uncovering the associated modes of thinking from without. But building theories of aspects of mind and language using such notions as part of the theoretical vocubulary risks begging questions. To the extent that such notions appear in theoretical inquiries directed at human beings and their capacities, caution is required. In many cases the use of these notions mark places where further inquiry remains to be carried out, not places where real explanation occurs.

Even if future semantic theories take the sort of shape familiar from truth-conditional model-theoretic semantics, it may be that the models that we ultimately need to postulate have an intricate structure, and one reflective in various ways of human modes of thinking. If so, one can expect there to be difficult questions about how exactly the constraints on models that sentences express—truth-conditions,

in a technical sense—relate to the way the world is, the latter understood from a theoretical perspective, from the third-person perspective of scientific inquiry into mind and language.

Let me sketch the way the preceding interacts with one common perspective within formal semantics. In his introduction to Montague's collected papers, Thomason (1974) writes:

> *we should not expect a semantic theory to furnish an account of how any two expressions belonging to the same syntactic category differ in meaning* . . . 'Walk' and 'run', for instance, and 'unicorn' and 'zebra' certainly do differ in meaning, and we require a dictionary of English to tell us how. But the making of a dictionary demands considerable knowledge of the world [of a sort the semantic theorist should not be expected to furnish]. (48; italics his)[12]

Something like this seems to me a prevailing sentiment among semanticists. A version of the thought is as follows: semantic theory captures an abstract level of meaning-relevant structure, but at a certain point, we may reach distinctions in meaning which are, from the point of view of semantic theory, simply brute (as might be the case with 'walk' and 'run', or 'unicorn' and 'zebra'). If further illumination is to be had for these differences of meaning, it is to be sought outside of semantic theory proper.

It is important to note that on this view of the domain of semantic theory, many differences in meaning that would in fact be critical to what a sentence communicates—to 'what is says', understood in any sort of common-sensical way (for instance, the sense operative in Grice 1975) are in a core respect outside of its explanatory domain. Semantics captures systems of meaning relationships and influences between expressions,[13] but typically to a degree that substantially underdetermines communicative import. From a semantics-internal perspective (and supposing Thomason is right about 'zebra' and 'unicorn'), the difference in 'truth-conditions' between 'Bill is riding a zebra' and 'Bill is riding a unicorn' is a thin one—not thin in the radical sense of the deflationist, but still not thick enough to reflect meaning differences between 'zebra'

[12] Thomason later walks some of this back, adding:

> A policy of refusing to allow semantic distinctions between basic expressions would be harmful if followed uniformly. For instance, if linguistic evidence shows that 'every' and 'some' are best treated alike syntactically, this should not force us to exclude an account of university and existential quantification from our semantics. (49)

[13] I take it a main test—not the only one—for whether a semantic feature of an expression is to be handled within semantic theory proper is given by the question whether that feature is subject to any sort of systematic compositional influence. If it is, it probably belongs to semantics. What is 'systematic compositional influence'? Think of quantifiers binding pronouns, modals shifting the world of evaluation for clauses they embed, degree modifiers interacting with the scale structure of adjectives, prepositions interacting with aspectual structure, connectives interacting with pairs of elements from a uniform category, and so on; think perhaps also of presupposition projection, and of scalar implicature in complex sentences. Semantics, on this conception, is trained primarily on systematic meaning interactions. (Presumably the question whether any given feature is subject to any sort of systematic compositional influence is itself a theoretical one. It might require advances in semantic theory for the relevant feature to even come into focus. We shouldn't expect *a priori* general criteria for drawing the boundary between which features seeming to do with meaning get handled with semantics and which are to be dealt with elsewhere.)

and 'unicorn' which go much farther beyond the mere fact that their intensions are disjoint. The truth-conditions compositionally associated with these sentences is thus something more skeletal and indeterminate than it may at first appear. What this draws out is not an objection to semantic theory, but the degree to which it is sharply focused on select aspects of the range of phenomena informally classified under the heading 'meaning', and the degree to which it pulls apart from anything that could be called a theory of human communication. Frege's microscope metaphor comes to mind: in scientific inquiry we make progress by severely narrowing our focus.

There are different conceptions of what is supposed to happen beyond the lexical boundary where narrowly compositional semantic explanation putatively comes to an end. These differences tend to reflect different attitudes about how semantic theory fits in a broader theoretical picture. Linguists are apt at this point to start talking about 'lexical semantics'—for instance, Charles Fillmore's classic work, Cruse (1986), the work of the MIT lexicon project, Pustejovsky (1991); perhaps related work in the psychology of lexical acquisition (e.g., Bloom (2002) and references therein). Philosophers are more likely to emphasize a distinction between 'foundational' and 'descriptive' semantic projects (in the terminology of Stalnaker 1997). On a prevailing conception in this vein, it's not that compositional semantics traffics in 'skeletal' truth-conditions; it's just that the full explanatory story turns partly on issues of a foundational kind, issues not addressed in compositional semantics proper. Compositional semantics—the 'descriptive' semantic project—hypothesizes a lexicon which maps basic expressions to semantic values, and theorizes about how these semantic values systematically combine to form the semantic values of complex expressions. A formal statement of the lexicon is assumed to reuse, or effectively reuse, the object language in the metalanguage, at least for a wide swath of open class expressions: perhaps it maps 'London' to London, 'zebra' to (something tantamount to) the property of being a zebra, and so on. What then remains to be addressed is the question *in virtue of what* the lexical items have the semantic values that they have: granted that 'London' denotes London and 'zebra' denotes zebras (zebrahood, etc.), in virtue of what is this the case? This is taken to be handled by a separate but connected foundational semantic project—a project often assumed to be reductive in character, to be distinctively philosophical, and to be a core chapter of a theory of intentionality. Perhaps Kripkean causal chains are invoked (Kripke 1972/1980); perhaps information-theoretic connections between the agent and the environment (Dretske 1981; Stalnaker 1984); perhaps some kind of Fodorian asymmetric dependence relations (Fodor 1990a,b); perhaps Lewisian convention and eligibility (Lewis 1975, 1994); perhaps some admixure of these or something else.

This descriptive/foundational distinction has an attractive simplicity that can give it the appearance of inevitability, but I think this appearance belies a complex reality, a reality the distinction as traditionally understood distorts. I have already inveighed in some detail against the idea that the semantic values of expressions can be safely stated by using those very expressions in the metalanguage—an idea that typical

statements of the descriptive/foundational distinction presuppose. To be clear, the complaint is not one about the very idea of associating expressions with elements of an extra-mental reality. It isn't being denied that we can talk about the world. Rather, the problem is with adopting insufficiently subtle understanding of the semantic metalanguage. A second problem is that work in the foundational semantic vein rarely takes care to distinguish the linguistic notion of compositional semantic value from the philosophical notion of mental content. In many cases it is some notion of mental content, not the technical linguistic concept of semantic value, that theorists are foremost attempting to explicate and ground. (See Yalcin 2014 and citations therein for further discussion.) A third problem is that insofar as 'foundational semantics' has a clear agenda, it would be constrained by 'descriptive semantics' inasmuch as the latter states the semantic value facts that the former project is charged with grounding. But the semantic values facts postulated in modern natural language semantics are continually shifting, undergoing constant revision as theorizing proceeds. Since its boundaries are not clear, neither is the agenda of foundational semantics. One might worry that foundational semantics is animated by a curious emphasis on reduction, one that is premature in the present state of understanding.

If current semantics seems to fail to provide much in the way of explanation for how it is that (e.g.) 'walk' and 'run' differ in meaning, that may well be more a reflection of our parochial location in the history of the development of the science than a fact about the intrinsic explanatory limits of compositional semantics. In any case, this does not seem to be something we can safely judge in advance. It certainly seems that more and more has fallen into the realm of compositional semantic explanation as theorizing has progressed—think, for instance, of the aspects of meaning to do with event structure, aspectual structure, and scale structure, or the way that scalar implicature, originally postulated as an extra-semantic phenomenon, is now thought by many to interact with compositional semantics in nontrivial ways.

It is one thing for the semantic theorist to say: 'We're not going to deal in any deep way with the semantic difference between "walk" and "run" (etc.) right now; we're finding it more fruitful to focus on other things.' This is an unobjectionable methodological posture, akin to looking for the keys where the light happens to be. It would be another thing to declare that semantic distinctions like those between 'walk' and 'run' will never fall within the explanatory scope of natural language semantics, and are officially out of its purview. This seems like mere speculation about the future of semantic theory. It risks mistaking the present limits of our understanding with the joints of semantic nature.

The lexicon has often been thought to be the domain of the idiosyncratic. It certainly can seem so. But then, such a view about syntax would have seemed extremely natural eighty years ago. As in the case of grammar, our intricate lexical knowledge needs to get squared with the facts of acquisition, creating pressure to find underlying systematicity. How continuous will this underlying systematicity be with the sort of systematicity natural language semantics already teaches us about? Perhaps

it will quite discontinuous, in ways that justify the view that lexical semantics is its own thing. But perhaps there will be nontrivial continuity in important respects. It is not clear there is anything to say here, except: let's see what happens.

12.5 Natural Language Semantics as Model-Based Science

We might achieve some limited clarity on the above matters if we understand natural language semantics to be substantially a model-based science, again understanding 'model' not in the technical sense of model theory (the sense operative in 'model-theoretic semantics') but rather in the general sense of a scientific model, a notion discussed in the philosophy of science from various points of view by, e.g., Suppes (1961); Gibbard and Varian (1978); Giere (1988); Downes (1992); Morgan and Morrison (1999); Godfrey-Smith (2006); Weisberg (2007); Frigg and Hartmann (2012); among many others.[14] The point to stress is that model-building is a strategy of indirect representation. Models are 'mediating instruments' in Morrison and Morgan (1999)'s phrase. They are representations of a distinctive sort. They are idealized structures we use to represent select aspects of the world we are seeking some theoretical understanding of. I take it that in natural language semantics, the aspect of reality we are seeking some understanding of is a dimension of human linguistic competence—informally, knowledge of meaning. Competent speakers of a language know ('cognize', etc.) the meaning features of expressions of their language. The semanticist is interested in modeling this state of mind and the associated semantic features. Elements of models in semantics—often constructed with tools from the mathematical theory of models—are used to represent these semantic features. We can think of the model associated with a given model-theoretic truth-conditional semantics for some fragment of language, for instance, as modeling the relevant aspect of speaker competence—a postulated mental state. It is a model of the relevant state of mind—informally, that of knowing the meaning of the language fragment.[15]

[14] Barwise briefly suggests something like this gloss in Barwise (1987). Considering the question 'How are logicians to think about the relation between the artificial languages they invent and the languages people actually use?', he suggests that in many cases, this kind of enterprise should be understood as model-building. Montague's intensional logic, for instance, gives a model of natural language in the way that 'a globe is a model of the earth' (4). (Barwise also suggests that giving a model of some phenomena is separable from giving a theory of it, and he suggests situation semantics is a theory, rather than a model. It seems to me a mistake, however, to think of model-building and theory-building as separate enterprises. Theories are typically stated in terms of features of models, and models are typically designed to faciliate the statement of theoretical generalizations.)

[15] 'Informally' since it is inadvisable to understand the relevant notion of semantic competence as a propositional attitude, much less as literally a state of knowing. This point is discussed further in Yalcin (2014).

Note, I am not saying that the models constructed by semanticists are 'represented' by mental states of semantic competence. That would be confused, and analogous to saying that DNA 'represents' the double-helix model or atoms 'represent' the Bohr model.

As with any model-based scientific inquiry, questions arise about how the model constructed should be understood to map on to the 'target system'—the part of reality—being modeled. That issue is often fraught and subject to continual reassessment as theorizing proceeds. The matter can be complicated by the fact that it may not be clear that there is a way to talk about the 'target system' with any precision without the help of the model itself. Questions of how best to think about the relationship between a scientific model and the aspects of reality it models are often tied to questions of reduction, particularly to questions about whether the entities countenanced in one model are systematically reducible to the entities of some other model at an appropriate 'lower level' of description. In these cases, progress may be impeded by lack of any clear sense what the relevant lower level of description is. Further, as Godfrey-Smith (2006) observes generally about models in science, 'two scientists might use same model for the same target system, but with different resemblance relations [between the model and the target system] in mind' and 'There is usually something more like a multidimensional space of different ways in which a model system might resemble a target' (733).[16] There is no particular reason to doubt that these points hold true for the models constructed in natural language semantics as much as they hold for models constructed elsewhere. In any case, one does not need to settle these issues in advance to locate natural language semantics as a model-based science; such questions do not arise for semantics uniquely.

This take on semantics differs from what is perhaps the standard picture, the picture on which what natural language semantics does is relate expressions 'directly' to objects and properties in extra-linguistic reality (and not to entities in an intermediate kind of theoretical representation, as I am suggesting). I have tried to indicate in the preceding pages what is problematic about that gloss. The metalanguage of semantics, I am suggesting, is language for articulating features of the theorist's model. The interpretation function given in such a model associates pieces of language with semantic features, the latter usually modeled via certain elements of model structures in the logician's sense. Understanding semantics as model-based in this way, it is not really appropriate to say that the interpretation function maps expressions to what they 'refer' to, if 'refer' is used in the normal, everyday sense. This is a category mistake, though one easy to miss. Generally speaking, a theoretical model of some aspects of

[16] Godfrey-Smith also writes:

> When much day-to-day discussion is about model systems, disagreement about the nature of a target system is less able to impede communication. The model acts as a 'buffer,' enabling communication and cooperative work across scientists who have different commitments about the target system. (Godfrey-Smith 2006: 739)

This seems accurate as a description of practice in natural language semantics. There is no shortage of diversity of views about how precisely to frame what it is that semantic theory models—how to conceive of the 'target system' of the theory. Despite the potential for foundational differences at this level, there is rather substantial agreement among theorists about how to go about building and comparing formal semantic theories.

reality will not involve the claim that those aspects somehow 'refer' to elements of the model constructed. If we use the billiard ball model to model the behavior of a gas, for instance, we are not involved in the strange hypothesis that the thing being modeled—the gas—'refers' to a collection of billiard balls in the model. There is obviously a correspondence intended, but the relationship is not one of reference. Gases do not 'refer' to explanatory abstractions dreamt up by scientists. Rather, the model is a representation we use to draw out the properties of gases we are interested in understanding, partly by idealizing away other features. Is there any special reason to treat the models built in natural language semantics differently here? In semantics we are modeling language in its semantic respects. The additional claim that the things being modeled—items of language, or our knowledge of thereof—'refer' to the items in the models we invent seems about as otiose and misguided as the corresponding claim about gases and our models of them.

There is a well-motivated theory-internal concept of compositional semantic value. This notion is not the ordinary everyday notion of reference. The use of a nontechnical notion of reference is not required in semantics. (Nor does semantic theory impugn everyday platitudes about reference. Semantic theory is not a theory of those platitudes.)[17]

There is little consensus in the philosophy of science about what models are, and about how best to situate the role of models and model-building in scientific inquiry generally.[18] Difficult questions arise here, some already noted above. My main interest has been to start the framing of natural language semantics as (substantially) model-based, not to resolve the basic questions about how the strategy of model-based science works in general. From the perspective developed here, the traditional foundational semantic question 'In virtue of what does an expression have the semantic value that it has?' transforms into a question more like: 'In virtue of what is a given semantic model of a language fragment a good one?'—a question which is a special case of a more general one about models in science ('In virtue of what is a given scientific model of some phenomena a good one?'). The 'mediating instrument' role of semantic models—their status as scientific representations—is important to acknowledge in any discussion of the foundations of semantics.

[17] There is a tradition of trying to approach scientific modeling by construing scientific theories syntactically, understanding a theory as a deductively closed set of sentences of some formal language; and we might then try to understand all model-building in science as effectively enterprises that supply theories with models in the technical, model-theoretic sense of 'model'. (This is sometimes called the 'semantic view' of scientific models; see, e.g., Suppes 1961.) I am presupposing without argument that this is the wrong way of thinking about modeling in science. For discussion of the problems with this approach to theories and to scientific models, see Downes 1992; Godfrey-Smith 2006; Frigg and Hartmann 2012.

[18] A recent extensive overview article on the topic concludes noting that 'despite the fact that they have generated considerable interest among philosophers, there remain significant lacunas in our understanding of what models are and of how they work' (Frigg and Hartmann 2012). For a recent overview on the metaphysics of models, see Gelfert (forthcoming).

The human capacity to speak and understand language and the human capacity to construct explanatory models are superficially alike in that they are, broadly speaking, representational capacities, and capacities unique to our species. Insofar as possible, we should like accounts of each of these capacities. But we should not assume in advance the accounts will be alike. We should take care to separate the questions:

- What does a good model of human semantic competence with a given language look like?
- How do human beings use models to understand and explain features of reality?

I have been construing natural language semantics as especially occupied with the first question. It seems to me that the modern history of semantics is a history of impressive progress on this question. The second question is to do with psychology of scientific theorizing, but also with the large-scale philosophical question how to conceive of the relation between a scientific model and the reality it models.

Failure to separate these questions can lead to confusion. Some philosophers—Davidson, for instance—seem to operate under the assumption that natural language is in some sense the fundamental medium of articulate human thought, undergirding even the most abstract scientific thinking. Technical scientific language, insofar as it is intelligible, is really a sort of sophisticated shorthand, and always admits of a (perhaps complicated) natural language gloss. These philosophers think that if anything can be said clearly, it can be said in plain English. I have been taking it for granted that the natural sciences supply myriad counterexamples to this view. As Frege and Tarski both emphasized, sometimes the only way to say things clearly is to depart from natural language. Maybe Wittgenstein was right to say that what can be said at all can be said clearly, but from this it doesn't follow that it can be said in a natural language. When we depart from natural language and resort to technical language and abstract models in an effort to develop a scientific conception of aspects of the world, to what extent are we then deploying mental representations or modes of thinking of a character very different from natural language? On the face of it, it seems hard to believe that there should be any presumption that our accounts of these two representational tendencies—that of speaking and understanding a natural language, and that of deploying abstract formal representations like models in the service of scientific theory-building—will look very similar.

This point matters to the foundations of semantics for various reasons. One is that there is a vague temptation to think that natural language semantics faces a unique sort of methodological challenge. Isn't semantic theory unable to achieve a theoretical vantage point completely free of assumptions about the meaningfulness of language, owing to the simple fact that any theory must be stated using language or symbols whose meanings are taken for granted by, rather than explained by, the theory? Doesn't this therefore mean that semantic theory is subject to a sort of fundamental incompleteness, or a problematic explanatory circularity? How can semantic theory paint the spot it is standing on? This suspicion really isn't clear enough to take

seriously, but a worry along such lines might lead one to suspect that semantics calls for some kind of nonstandard methodology. I think the appearance of a problem here is engendered by the mistaken idea that natural language semantics itself somehow takes place entirely in natural language. Once we are in the realm of denotation functions, models, lambdas, and so on, we are using representational devices and associated modes of thinking that are not plausibly part of natural language or natural language understanding, and which are not plausibly understood as extensions of natural language.[19] There isn't eventually some moment where we put lambda terms (say) themselves in interpretation brackets, trying to turn semantic theory on its own technical materials. This should diminish the concern that the explanatory tools we are using are somehow themselves part of the system being explained in a potentially question-begging way. The semanticist can paint the whole floor because she can stand in another room and reach in.

Sometimes appeal to technical language is used to evade philosophical problems by kicking up dust. But sometimes the philosophical problem really owes to a misunderstanding about the role of the technical language itself.[20]

References

Bartsch, Renate and Theo Vennemann (1972). The grammar of relative adjectives and comparison. *Linguistische Berichte*, **20**: 19–32.

Barwise, Jon (1987). Noun phrases, generalized quantifiers and anaphora. In Peter Gärdenfor (ed.), *Generalized Quantifiers*. Dordrecht: D. Reidel Publishing Company, pp. 1–29.

Bloom, Paul (2002). *How Children Learn the Meanings of Words*. Cambridge, MA: The MIT Press.

Chomsky, Noam (1959). A review of B.F. Skinner's *Verbal Behavior*. *Language* **35**(1): 26–58.

Chomsky, Noam (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Cruse, D. Alan (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Davidson, Donald (1967a). Truth and meaning. *Synthese* **17**: 304–323. (Reprinted in Davidson 1984: 17–36).

Davidson, Donald (1967b). The logical form of action sentences. In Nicholas Rescher (ed.), *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press, pp. 81–95. (Republished in Donald Davidson, *Essays on Actions and Events*, Oxford University Press, Oxford, 1980).

Davidson, Donald (1973). In defense of convention T. In Hugues Leblanc (ed.), *Truth, Syntax and Modality*. Amsterdam: North-Holland, pp. 76–86. (Reprinted in Davidson 1984: 65–75).

[19] It isn't that semantic theorists have somehow devised a new dialect of natural language, one that their children might then naturally acquire.

[20] For helpful comments, I am indebted to Nate Charlow, Daniel Lassiter, Stephanie Leary, David Plunkett, Adam Simon, Brian Rabern, and Daniel Rothschild, and especially to Wesley Holliday and Wolfgang Schwartz.

Davidson, Donald (1977). Reality without reference. *Dialectica* **31**(3–4): 247–58. (Reprinted in Davidson 1984: 215–225).

Davidson, Donald (1984). *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, Oxford.

Davidson, Donald (1990). The structure and content of truth. *Journal of Philosophy* **87**(6): 279–328.

Downes, Stephen M. (1992). The importance of models in theorizing: A deflationary semantic view. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. JSTOR, pp. 142–153.

Dowty, David R. Robert E. Wall, and Stanley Peters (1980). *Introduction to Montague Semantics*. Dordrecht: D. Reidel Publishing Company.

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

Dummett, Michael (1973). *Frege: Philosophy of Language*. Harvard University Press, Cambridge, MA, 2nd edition.

Fodor, Jerry (1990a). A theory of content, I: The problem. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, Jerry (1990b). A theory of content, II: The theory. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Foster, J. A. (1976). Meaning and truth theory. In Gareth Evans and John McDowell (eds.), *Truth and Meaning: Essays in Semantics*. Oxford: Oxford University Press, pp. 1–32.

Frigg, Roman and Stephan Hartmann (2012). Models in science. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, fall edition.

Gelfert, Axel (forthcoming). The ontology of models. In Lorenzo Magnani and Tommaso Bertolotti (eds.), *Springer Handbook of Model-Based Science*. London: Springer.

Gibbard, Allan and Hal R. Varian (1978). Economic models. *The Journal of Philosophy* **75**(11): 664–77.

Giere, Ronald (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Glanzberg, Michael (2014). Explanation and partiality in semantic theory. In Alexis Burgess and Brett Sherman (eds.), *Metasemantics: New Essays on the Foundations of Meaning*. Oxford: Oxford University Press, pp. 259–92.

Godfrey-Smith, Peter (2006). The strategy of model-based science. *Biology and Philosophy* **21** (5): 725–40.

Grice, H. Paul (1975). Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Volume 3: Speech Acts*. London: Academic Press, pp. 41–58.

Heim, Irene (1982). The semantics of definite and indefinite noun phrases. PhD thesis, University of Massachusetts.

Heim, Irene and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Jacobson, Pauline (2014). *Compositional Semantics: An Introduction to the Syntax/Semantics Interface*. Oxford: Oxford University Press.

Kamp, Hans (1981). A theory of truth and semantic representation. In Jeroen A. Groenendijk, Theo Janssen, and Martin Stokhof (eds.), *Formal Methods in the Study of Language*. Amsterdam: Mathematisch Centrum, University of Amsterdam, pp. 277–322.

Kaplan, David (1977/1989). Demonstratives. In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press, pp. 481–563.

Kennedy, Chris (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* **30**(1): 1–45.

Keshet, Ezra (2008). Only the strong: Restricting situation variables. *Proceedings of SALT* **18** 483–95.

Keshet, Ezra (2010). Split intensionality: A new scope theory of de re and de dicto. *Linguistics and Philosophy* **33**(4): 251–83.

Kripke, Saul (1972/1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Larson, Richard and Gabriel Segal (1995). *Knowledge of Meaning*. Cambridge, MA: MIT Press.

Lewis, David K. (1975). Languages and language. In Keith Gunderson (ed.), *Language, Mind, and Knowledge. Minnesota Studies in the Philosophy of Science*, vol. 7 Minneapolis, MN: University of Minnesota Press pp. 3–35.

Lewis, David K. (1980). Index, context, and content. In S. Kanger and S. Ohman (eds.), *Philosophy and Grammar*. Dordrecht: D. Reidel, pp 79–100.

Lewis, David K. (1994). Reduction of mind. In S. Guttenplan, (ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell, pp. 412–31.

Link, Godehard (1983). The logical analysis of plurals and mass terms: A lattice-theoretic approach. In Rainer Bauerle, Christoph Schwarze, and Arnim von Stechow (eds.), *Meaning, Use and Interpretation of Language*. Berlin: Walter de Gruyter, pp. 303–323.

Morgan, Mary and Margaret Morrison (eds.) (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Morrison, Margaret and Mary Morgan (1999). Models as mediating instruments. In Margaret Morrison and Mary Morgan (eds.), *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press, pp. 10–37.

Ninan, Dilip (2010). Semantics and the objects of assertion. *Linguistics and Philosophy* **33**(5): 355–80.

Pustejovsky, James (1991). The generative lexicon. *Computational linguistics* **17**(4): 409–41.

Rabern, Brian (2012). Against the identification of assertoric content with compositional value. *Synthese* **189**(1): 75–96.

Rabern, Brian (2013). Monsters in Kaplan's logic of demonstratives. *Philosophical Studies* **164** (2): 393–404.

Rosen, Gideon (2010). Metaphysical dependence: Grounding and reduction. In Bob Hale and Aviv Hoffmann (eds.), *Modality: Metaphysics, Logic, and Epistemology*. Oxford: Oxford University Press, pp. 109–36.

Schwarz, Florian (2012). Situation pronouns in determiner phrases. *Natural Language Semantics* **20**(4): 431–75.

Stalnaker, Robert (1984). *Inquiry*. Cambridge, MA: MIT Press.

Stalnaker, Robert (1997). Reference and necessity. In B. Hale and C. Wright (eds.), *A Companion to the Philosophy of Language*. Oxford: Blackwell, 534–54.

Suppes, Patrick (1961). A comparison of the meaning and uses of models in mathematics and the empirical sciences. *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences*. London: Springer, 163–77.

Tarski, Alfred (1936). Der Wahrheitsbegriff in den formalizierten Sprachen. *Studia Philosophica* **1**: 261–405.

Tarski, Alfred (1944). The semantic conception of truth: And the foundations of semantics. *Philosophy and Phenomenological Researc*, **4**(3): 341–76.

Tarski, Alfred (1956). The concept of truth in formalized languages. In Alfred Tarski (ed.), *Logic, Semantics, Metamathematics*. Oxford: Oxford University Press, pp. 152–278. (Translated by J.H. Woodger).

Thomason, Richmond (1974). Introduction. In Richmond Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, CN: Yale University Press, pp. 1–69.

Weisberg, Michael (2007). Who is a modeler? *The British Journal for the Philosophy of Science* **58**(2): 207–33.

Yalcin, Seth (2014). Semantics and metasemantics in the context of generative grammar. In Alexis Burgess and Brett Sherman (eds.), *Metasemantics*. Oxford: Oxford University Press, pp. 17–54.

Yli-Vakkuri, Juhani (2013). Propositions and compositionality. *Philosophical Perspectives* **27**(1): 526–63.