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 first critique some standard ways of thinking about how the metalanguage of semantics is to be understood.

1 Heim and Kratzer on showing meanings

At the beginning of their excellent and influential textbook in natural language semantics (Heim and Kratzer [1998]), Heim and Kratzer 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:

\[
\begin{align*}
[\text{Ann}] &= \text{Ann} \\
[\text{smokes}] &= \lambda x. x \text{ smokes}
\end{align*}
\]

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:

*For Brian Rabern and Derek Ball, editors, The Science of Meaning, Oxford University Press, forthcoming.*
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:

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

And then we might have calculated as follows:

\[
[\text{Ann smokes}] = [\text{smokes}](\text{Ann}) = \lambda x. x \in \{\text{Ann}, \text{Jan}\} (\text{Ann}) = 1 \text{ iff Ann smokes}
\]

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 expression 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. (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:

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

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:

\[
\text{[every]} = \lambda f_{et}. \lambda g_{et}. \text{every } x \text{ such that } f(x) = 1 \text{ is such that } g(x) = 1
\]
\[
\text{[every]} = \lambda_x f_x \cdot \lambda_x g_x. \text{for all } x, \text{ if } f(x) = 1 \text{ then } g(x) = 1
\]
\[
\text{[every]} = \lambda_x f_x \cdot \lambda_x g_x. \text{for every } x, \text{ if } f(x) = 0 \text{ or } g(x) = 1
\]
\[
\text{[every]} = \lambda_x f_x \cdot \lambda_x g_x. \text{for all } x, \text{ if } g(x) = 1 \text{ or } f(x) = 0
\]

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

\[ \text{smokes} = \lambda x. x \text{ smokes} \]

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

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

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.

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 paper). The notion of the truth-condition of a sentence is a fundamentally modal notion: it concerns what would be were the sentence true. A “condition”

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.)
demuded 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.

2 Disquotation

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

\[[\text{Ann smokes}] = 1 \text{ iff } \text{Ann smokes}\]

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 as 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 _______.

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 meta-theoretical reflections—he is the barely-concealed wizard behind the curtain. 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

---

2It 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.
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. 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

---

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

4We 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
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], Tarski [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 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].

\footnote{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].}

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 \textit{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 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):

\[
[\text{smokes}] = \lambda x. x \text{ smokes}
\]

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:

\[
\text{Val}(x, \text{ponders}) \text{ iff } x \text{ ponders}
\]

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.\footnote{Many philosophers of language undergo years of radical interpretation therapy in order to unsee this problem.}

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 \emph{not} depending on competence with ordinary language or ordinary concepts. And the righthand sides displayed above are really considerably more remote in their content from anything expressible in ordinary language than Heim and Kratzer, or Larson and Segal, make it appear.

\section{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.\footnote{Glanzberg \citeyear{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.} 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 inexacty 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 individual
als 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.

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 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.

---

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.

10 In this section of the paper I seem to differ with Glanzberg [2014], who holds that disquo-
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:

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

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.\(^\text{11}\) 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 about what the homonymous ‘smokes’ of the metalanguage above means, or how it works. So these homonyms are not synonyms.

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

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

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, commonsense 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 commonsense. (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 vocabulary 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 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.

5 Natural language semantics as model-based science

We can 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. 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 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.\textsuperscript{12}

As with any model-based scientific inquiry, questions arise about how the model constructed should be understood to map on to the ‘target system’ being modeled (as emphasized the last paragraph of the preceding section). 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 \textsuperscript{[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).\textsuperscript{13} 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

\textsuperscript{12}“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 \textsuperscript{[2014]}.

\textsuperscript{13}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 \textsuperscript{[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.
‘directly’ to objects and properties in extra-linguistic reality (and not to entities in some sort of 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 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 atoms “refer” to the billiard balls of our model. Or again, if we use the Lotka-Volterra equations to model the population dynamics for the interaction of two species standing in a predator-prey relation, we are not saying that the populations, or members of the population, “refer” to features of the model.\footnote{There is no special reason to treat the models built in natural language semantics differently here. In semantics we are interested in modeling reality in its semantic aspects. Engaging in that project does not require the use of a nontechnical notion of reference. (Nor does not impugn everyday platitudes about reference. Semantic theory is not a theory of those platitudes.)} There is no special reason to treat the models built in natural language semantics differently here. In semantics we are interested in modeling reality in its semantic aspects. Engaging in that project does not require the use of a nontechnical notion of reference. (Nor does not impugn everyday platitudes about reference. Semantic theory is not a theory of those platitudes.)

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.\footnote{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.} 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, many questions traditional to the foundations of semantics (e.g., ‘In virtue of what does an expression have the semantic value...'}
that it has?') will connect with philosophical questions about models generally (e.g., ‘In virtue of what is a given scientific model of some phenomena a good one?’).

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 expect the accounts to be different. 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. 16

References


16For helpful comments, I am indebted to Nate Charlow, Daniel Lassiter, Stephanie Leary, David Plunkett, and Adam Simon, and especially to Wesley Holliday and Wolfgang Schwartz.


