Notes on Implicit Definitions and Truth-Value Semantics: Commentary on Hans Halvorson's *The Logic in Philosophy of Science* (APA 2020 Symposium)

Holger Andreas

University of British Columbia (Okanagan)

Contents

1 Praise and Endorsement

It is a great pleasure and honour for me to discuss Hans Halvorson's book on the logic in philosophy of science. Thank you Samuel, thank you Hans!

It is uncontroversial, I think, to say that modern logic of science starts with Carnap's *Aufbau* and his *Logical Syntax of Language*. Not surprisingly, Hans' book is very much in this Carnapian tradition. He himself says that he wants to take Carnap's syntactic research programme to 'a higher level of nuance and sophistication' (p. 11).

Having dedicated my PhD thesis to Carnap's logic of science, I was excited to see this continuation of Carnap's work. Let me briefly try to explain what seems to be the genuinely novel and most significant contribution of Hans' book. To the best of my knowledge, it is the first monograph that applies *category theory* – in a comprehensive and sustained manner – to problems in the philosophy and logic of science. The book begins with a self-contained exposition of category theory, and then expounds the semantics of propositional and first-order logic in a category-theoretic fashion. Finally, the category-theoretic point of view is exploited to address profound philosophical problems, most notably the problem of theoretical equivalence and Putnam's model-theoretic argument against (metaphysical) realism. Outstanding technical results are:

- An account of what it means that the *categories* of two scientific theories are equivalent.
- A novel criterion of intertranslatability between theories, called *Morita equivalence*, which is more liberal than definitional equivalence.
- The demonstration that Morita equivalence implies categorical equivalence (between theories).

To describe Hans' book and its overall objective in two sentences: it is comparable to the *Architectonic of Science* by Balzer, Moulines, and Sneed. As the latter tried to show that Bourbaki's semiformal, set-theoretic reconstruction of mathematics can be carried over to scientific theories, so does Hans show how we can fruitfully use category-theoretic notions in an analysis of scientific theories.

Now, I'd like make some comments and observations, all of which are at least as constructive as they are critical. The first observation is from the perspective of Carnap and Ramsey, the second from the perspective of Putnam. Finally, I'd like to ask Hans a simple question from the perspective of alternatives to standard semantics of first-order logic.

2 Weak and Strong Implicit Definitions

As is well known, the notion of an implicit definition is, at least, loosely tied to Hilbert's work on the foundations of geometry. In his *General Theory of Knowledge* (Allgemeine Erkenntnislehre), Moritz Schlick used this notion to describe the import of Hilbert's work. Since then, the notion of an implicit definition has been used in various ways.

In his book, Hans takes a closer look at an attempt to clarify functionalism in the philosophy of mind, according to which mental properties (of folk psychology) may be *implicitly defined* by physical properties. Functionalism aims to expound a non-reductive physicalism. It is non-reductive because it denies that mental properties are explicitly definable in terms of physical properties. Rather, mental properties are 'defined in terms of the role they play vis-a-vis each other and the physical properties'. (p. 250) This idea gave rise to *Ramsey Sentence Functionalism*, depicted by Hans along the following lines.

- T: theory that implicitly defines the mental properties.
- Σ: vocabulary of the physical properties.
- $r_1, \ldots, r_n$: symbols for the mental properties.

> T provides functional definitions of $r_1, \ldots, r_n$ in terms of Σ just in case, in each model M of the Ramsey sentence T^R, there are unique realizing properties $M(r_1), \ldots, M(r_n)$.
>
> It's easy to see then that T provides functional definitions of $r_1, ..., r_n$ in terms of Σ only if T implicitly defines $r_1, ..., r_n$ in terms of Σ. Indeed, if M and N are models of T, then $M|_\Sigma$ and $N|_\Sigma$ are models of T^R, and it follows from the uniqueness clause that $M(r_i) = N(r_i)$. It then follows from Beth's theorem that T explicitly defines $r_1, \ldots, r_n$ in terms of Σ. (p. 250)

I think Hans is a little bit quick here, and focuses too narrowly on a specific understanding of implicit definitions – the one used in Beth's definability theorem. So, I would like to suggest distinguishing between different notions of an implicit definition. To this end, let us view implicit definitions, in general, as *semantic constraints*: if the axioms of T implicitly define theoretical terms $\tau_1, \ldots, \tau_n$, then all (intended) interpretations of the theoretical terms must satisfy these axioms. Now, we can distinguish between different types of semantic constraints:

(1) Suppose we have an axiomatic theory T where none of the descriptive terms are considered to have an antecedent meaning or interpretation. Then, the

interpretation of the descriptive terms is only constrained by the condition that it must satisfy the axioms.

(In his correspondence with Frege, Hilbert quite explicitly affirms this view with regard to his *Foundations of Geometry*. In Hans' book, the view is aptly described as *de-interpretation*.)

(2) Some of the descriptive terms have an antecedent interpretation. Let's call these the *observation terms*, while the others are called *theoretical*. Then, the interpretation of the theoretical terms is constrained by two components: (i) the interpretation of the observation terms and (ii) the axioms of T. But these two constraints *need not* result in a unique interpretation of the theoretical terms. So, we must further distinguish:

   (a) The axioms of T *merely* constrain the interpretation of the theoretical terms, without uniquely determining their interpretation, even in the context of an antecedent interpretation of the observation terms. (I will argue that this was Carnap's and Ramsey's view of the semantics of theoretical terms.)

   (b) The axioms of T constrain the interpretation of the theoretical terms such that, for any interpretation of the observation terms, there is only one interpretation of the theoretical terms that satisfies this constraint. (This is how Lewis Lewis [9] thinks we should view the semantics of theoretical terms.)

If we differentiate between these types of semantic constraints, we can distinguish between different types of implicit definitions:

(1) Pure implicit definitions; no antecedent interpretation of any descriptive term.

(2) Impure implicit definitions that merely constrain the interpretation of the theoretical terms.

(3) Impure implicit definitions that uniquely determine the interpretation of the theoretical terms.

Obviously, only the third type of implicit definition satisfies the strict premises of Beth's definability theorem. Neither Ramsey nor Carnap, I think, had such a strict understanding of the semantics of theoretical terms in mind.

Historical note on Ramsey and Carnap. What is the motivation for Ramseyfication in Ramsey's seminal paper "Theories" [11]? Ramsey thinks that assertions made in the theoretical language are not 'propositions by themselves' (p. 231). The meaning of such assertions rather depends on the context of some theory T as well as the stock of other assertions made in the language of T. Since there was no formal semantics available at that time, he chose to express this semantic dependency using second-order logic. This leads to the following analysis of theoretical propositions ϕ (assertions made in the theoretical language):

$$\exists\overline{X}\ (\mathrm{Stock}(\overline{X}/\overline{\tau}) \wedge \phi(\overline{X}/\overline{\tau}))$$

- Stock: union of the axioms of the theory T and all propositions of the observation language presumed to be true.
- $\overline{X}$: sequence of second-order variables.
- $\overline{\tau}$: sequence of theoretical relation symbols.

At the same time, Ramsey emphasizes that both

$$\exists\overline{X}\ (\mathrm{Stock}(\overline{X}/\overline{\tau}) \wedge \phi(\overline{X}/\overline{\tau})) \tag{ϕ^R}$$

and

$$\exists\overline{X}\ (\mathrm{Stock}(\overline{X}/\overline{\tau}) \wedge \neg\phi(\overline{X}/\overline{\tau})) \tag{$(\neg\phi)^R$}$$

may well be true [11, p. 231n]. But this means that the two semantic constraints – the axioms of T and the given interpretation of the observation language – do not uniquely determine an interpretation of the theoretical language. So, the crucial premise of Beth's definability theorem is violated. To be more precise: this premise is violated if the 'stock' of propositions contains all true propositions of the observation language and all objects of the domain are designated by a closed term of the observation language.

Admittedly, Ramsey does not take these considerations to the level of a formal semantics of theoretical propositions, according to which we could say that ϕ is true iff some Ramseyfication of ϕ is true, or better: the Ramseyfication of ϕ is true, while the Ramseyfication of $\neg\phi$ is false. I will address this problem below.

Let's move on to Carnap. In his paper "Observational Language and Theoretical Language" (*Beobachtungssprache und theoretische Sprache*), Carnap [2, 4] gives some hints indicating that the axioms of T, together with an antecedent interpretation of observation terms, do not determine the theoretical terms uniquely. At the core of this paper is what has been termed later on the *Carnap sentence*:

$$TC^R \to TC$$

Carnap explains the meaning of this sentence as follows [4, p. 83]:

> In case the world is so constructed that any n-tuple at all of mathematical entities exists which satisfies TC then the T-terms are to be understood in such a way that the things designated by them form such an n-tuple.

An n-tuple, in this explanation, is simply a sequence of sets such that these sets figure as (model-theoretic) interpretations of the theoretical terms. The wording of the explanation suggests that there may well be more than one n-tuple of mathematical entities that satisfy the conjunction TC of axioms of the theory, in the context of an antecedent interpretation of the observation language. If so, the question arises which of those n-tuples that satisfy TC should we take as denotation of the theoretical terms?

We have to read another paper by Carnap in order to find out what Carnap thinks is the right answer to this question, viz., his paper on the use of Hilbert's epsilon operator in scientific theories [3]. Loosely speaking, his answer in [3] is: it doesn't matter which of the n-tuples that satisfy TC – in the context of the antecedent interpretation of the observation language – you take as actual interpretation of the theoretical terms. Any will do, you can chose an arbitrary one; but stick to your choice once you have chosen a specific n-tuple. Formally, this is expressed using Hilbert's epsilon operator:

$$\overline{\tau} := \epsilon_Z \exists X_1 \ldots \exists X_n (Z = \langle X_1, \ldots, X_n \rangle \wedge TC(X_1, \ldots, X_n, o_1, \ldots o_m))$$

Recall that the epsilon operator picks out an arbitrary object out of a set of objects that satisfy a certain property.

$$\varepsilon x \phi(x)$$

thus stands for some object x for which the proposition $\phi(x)$ is true.

Clearly, the rationale for using the epsilon operator is that the axioms merely constrain the interpretation of the theoretical terms. There is no need for this constrain to result in a unique interpretation of theoretical terms. So, the strict premises of Beth's definability may well be violated. In other words, if Carnap had thought the axioms implicitly define the theoretical terms – in the sense of a unique determination – there would be no use for the epsilon operator.

However, neither Ramsey nor Carnap ever devised a formal semantics of theoretical terms that tells us when a theoretical sentence is true. Here is a proposal for such a semantics, which can be motivated by van Frasssen's notion of a supervaluation in [12] and Hintikka's use of modal logic in an analysis of knowledge and belief in [6]. (I'll come back to supervenience below). The semantics centres around the notion of an *admissible interpretation* of a language that contains theoretical terms (cf. Andreas [1, Sect. 3]):

Explanation 1. Admissible interpretation
An interpretation $\mathcal{A}$ is called *admissible* iff it satisfies two semantic constraints:

(1) $\mathcal{A} \models TC$

(2) $\mathcal{A}$ extends the antecedent interpretation of the observation terms to an interpretation of the complete language, which contains observation and theoretical terms.

If there is no $\mathcal{A}$ that satisfies both constraints, then $\mathcal{A}$ is given by the set of interpretations $\mathcal{A}$ that satisfy the second constraint.

Condition (2) is equivalent to saying that the model-theoretic reduct of $\mathcal{A}$ to the observation language equals the antecedently given interpretation of this language.

Let $\mathbf{A}$ be the set of admissible interpretations thus defined, and let ϕ be a theoretical sentence.

(1) ϕ is true iff, for all $\mathcal{A} \in \mathbf{A}$, $\mathcal{A} \models \phi$

(2) ϕ is false iff, for all $\mathcal{A} \in \mathbf{A}$, $\mathcal{A} \not\models \phi$

(3) ϕ is indeterminate iff, there is $\mathcal{A}_1 \in \mathbf{A}$ such that $\mathcal{A}_1 \models \phi$, and $\mathcal{A}_2 \in \mathbf{A}$ such that $\mathcal{A}_1 \not\models \phi$.

In brief, theoretical truth behaves like super-truth in a supervaluationst setting. Moreover, we can also think of the set $\mathbf{A}$ as a set of possible worlds – conceived to be possible from the viewpoint of the theory T. Then, using the basic idea of

Hintikka's modal account of belief and knowledge, we say that we believe a theoretical proposition ϕ iff ϕ is true in all worlds conceived to be possible (from the viewpoint of T).

Side remark: Of course, I was very pleased to see Hans using similar modal ideas to describe belief and acceptance in the context of a scientific theory at the end of his book. (p. 281n)

Finally, let us come back to supervenience, now using the modal semantics of theoretical terms. What is a theoretical (mental) property? Two answers suggest themselves. A theoretical property τ_i is the intersection of the interpretations of τ_i defined by the members $\mathscr{A}$ in $\mathbf{A}$. Alternatively, we can say that an object c has the theoretical property τ_i iff $\tau_i(c)$ is true on the modal semantics of theoretical terms. c does not have the theoretical property τ_i iff $\tau_i(c)$ is not true on the modal semantics. Finally, c is indeterminate as regards the theoretical property τ_i iff ....

Then, clearly, the theoretical properties *supervene* on the observational properties, without the axioms actually defining the theoretical properties. The unique determination of the theoretical properties is a mere limiting case of the modal semantics, where the set $\mathbf{A}$ of admissible interpretations is a singleton. To be more precise: if we hold the axioms fixed (as we should do since they implicitly define the theoretical terms), then any difference in truth value at the theoretical level must be grounded in a difference of the observational properties (represented by the antecedent interpretation of the observation terms). So, at least at the level of formal semantics, Ramsey-Sentence-Functionalism is coherent. We may well have supervenience of the mental properties on the physical properties – in the style of Ramsey –, without a reduction of the mental properties to physical properties in the style of explicit definitions.

3 Putnam's Model-theoretic Argument

This is on Putnam's model-theoretic argument. Put in a nutshell, Hans's criticism of this argument goes as follows: if we accept Putnam's model-theoretic argument, we adopt an extremely liberal criterion of equivalence between scientific theories: any consistent theory is equivalent with any other theory. So, we have no reason to prefer one theory over another. And this is of course an absurd conclusion.

Admittedly, I have always considered Putnam's argument a masterpiece of pristine, good-old-fashioned analytic philosophy, understood as the enterprise to solve, or dissolve, profound philosophical problems using modern mathematical logic. So, let me try to defend the argument, or better: its presumed philosophical conclusion.

My defence will be simple. The argument is meant to refute a certain view of models and reality, a view that Hans himself rejects and criticizes for reasons that may not coincide with Putnam's. Then, Putnam seeks to delineate an alternative view, which is loosely inspired by the semantics of intuitionist logic and its generalization by Dummett. But the delineation of the alternative remains tentative, and may not be considered to belong to the model-theoretic argument. At least the core of this argument is entirely negative as regards its philosophical consequences: it aims to refute a type of realism that views (model-theoretic) models as existing "out there" independent of any descriptions (p. 481). If I understand Hans' work on formal semantics correctly, he has little desire to defend such a view of models and reality. And so, I see much more convergence between Hans and Putnam as regards the negative consequences of the argument.

In what follows, I will focus on the semantic lessons that we may or may not learn from Putnam's argument. As regards the narrow model-theoretic part of the argument, I understand Hans approves of Putnam's demonstration: if a given theory T (understood as a set of sentence) has a model M, then we can construct numerous other models M' of T. And so the question arises which of these models may be considered an *intended interpretation* of our language. In other words, how can we know what the referents of our linguistic expressions are if our presumed knowledge about the world has the form of a theory T (i.e., a set of sentences we affirm).

Putnam thinks there is no way for this to be achieved, at least within the confines of metaphysical realism, according to which the truth-conditions of our claims about the world are somehow "out there" and independent of our theories and our use of language. If this correct, then Putnam has established the following implication: if we assume a certain type of realist, truth-conditional semantics, then we are agnostic about the meanings and referents of our linguistic expressions. We couldn't understand our own language. In brief, realist truth-conditional semantics implies that language learning is impossible. Thereby, Putnam strengthens similar arguments by Dummett.[1]

Of course, it is absurd that we don't understand our language. I take it that Putnam

[1] See Wright [13, Ch. 1] for a detailed and systematic account of Dummett's acquisition argument against truth-conditional semantics, purporting to show that language learning is impossible in a truth-conditional setting.

finds the semantic conclusions of the model-theoretic argument just as absurd as Hans does. It is important – for Putnam – that the argument has absurd conclusions because it is meant to refute a certain premise accepted at the outset of the model-theoretic demonstration. The premise is that truth conditions of our claims are out there in the world, prior to our theories and our use of language. Hans himself criticizes and rejects a similar premise, the assumption of language-free mathematical structures:

> The picture we get from the language-free semantic view is that mathematical structures are out there in the world, and that they are either isomorphic to each other or they are not. Of course, that picture completely ignores the fact that isomorphisms are defined in terms of language or, to put it more accurately, that isomorphisms relate mappings $\mathbf{M} \to \mathbf{Sets}$ and $\mathbf{M} \to \mathbf{Sets}$, which have a common domain. (p. 262)

Furthermore, Hans says in regard to Putnam's argument:

> None of us has the metalinguistic point of view that would permit us to see a mismatch between language and world. (p. 265)

But isn't this bad enough for the the type of realism targeted at by Putnam's argument? Doesn't this mean that we don't know what the referents of our linguistic expressions are once we think of models as 'existing "out there" independent of any description'? (Putnam [10, p. 481]).

However, Hans does not want us to read him as supporting the model-theoretic argument:

> Now, I suspect that some people might think that I've simply affirmed Putnam's conclusion, i.e., that I have embraced internal realism. I can neither affirm nor deny that claim (largely because of unclarity in the meaning of "internal realism"). But I insist that if Putnam's argument works, then we have no reason to discriminate between (ideal) consistent theories, and we should adopt an absolutely radical left-wing account of theoretical equivalence. I, for one, am loath to think that good theories are so easy to find. (p. 265)

For clarification, I think it is helpful to distinguish between the negative consequences of Putnam's argument and the resolution Putnam proposes:

(1) Negative consequences: a certain type of realism – figuratively described as invoking a God's eye point of view – leads to semantic agnosticism. That is, we don't know what our linguistic expressions refer to.

(2) Proposed resolution: some form of 'liberalized intuitionism' that allows us retain classical mathematical reasoning (Putnam [10, p. 479–481]). This may involve:

- Nonrealist semantics: meaning is use. Understanding the meaning of an expression comes with with using the expression.
- Internal Realism: idealized assertibility conditions instead of truth conditions.

These two points ((1) and (2)) are independent from one another. Pointing out problems with internal realism does not resolve Putnam's paradox. If internal realism is too vague and not properly defined, what could be the alternative semantics? In the final section, I would like to discuss an alternative to standard first-order semantics in this regard.

4 Truth-Value Semantics

Tarski-style model-theoretic semantics presents us with the following paradox. On the face of it, this type of semantics shows us how we can understand the interpretation of a language. If so, the metalanguage stands in need of an interpretation as well. So, we must assume an infinite hierarchy of metalanguages. Bas van Fraassen just indicated why the notion of such an infinite hierarchy is paradoxical. Putnam pointed out that the predicament of the model-theoretic argument recurs at the level of the metalanguage [10, p. 482]. In Hans's discussion of Putnam's argument, Hans seems to invoke the notion of a model of a theory in the metalanguage, and so a metalanguage of the metalanguage is needed.[2]

Conjecture: certain alternatives to Tarski-style model-theoretic may help us develop a less paradoxical notion of a metalanguage. In what follows, I would like to

[2] If I understand correctly, the background theory T_0 is first used to directly describe the world (domain of interpretation of Mette's theory) on page 264, and then considered to have a model given by the 'world W' on the next page. So, T_0 is first a theory given in the metalanguage, and then considered to have models.

outline why this conjecture could merit further consideration, and then to ask Hans what he thinks about it.

Some alternatives to Tarski-style model-theoretic semantics:

(1) Substitutional semantics

(2) Truth-value semantics

(3) Hintikka sets.

I shall focus on truth-value semantics. As is well known, the semantics of propositional logic in terms of truth-value assignments to propositional variables is easier and simpler than Tarskian model-theoretic semantics of first-order logic. The life of students and instructors would be easier if we could generalize the semantics of truth-value assignments to the first-order case. Why is this not possible? From the perspective of model-theoretic semantics, the problem is that we may have objects in the domain of interpretation that are not referred to by any closed term in the object language.

However, there is an elegant way to solve this problem. Let L be a first-order language. Let L^+ be a *term extension* of L. That is, L^+ is obtained from L by adding further terms to L, where L^+ has countably many terms. By definition, L is a term extension of itself. Further let v be an assignment of truth values to the atomic sentences in L^+.

Definition 1. Logical truth
A sentence ϕ of L is logically true – in the truth-vale sense – iff ϕ is true on all truth-value assignments v for all term extensions L^+ of L.

Definition 2. Logical Entailment
Let Γ be a set of sentences in L and ϕ be such a sentence. Γ logically entails ϕ – in the truth-vale sense – iff, for all truth-value assignments v and all term extensions L^+ of L, if all members of Γ are true on v, then ϕ is true on v as well.

Notably, this type of truth-value semantics is equivalent to Tarskian model-theoretic semantics. That is, soundness and completeness can be shown for the same deductive systems of first-order logic. Even 2-nd order logic and set theory have been studied using truth-value semantics.[3]

[3] See Leblanc [7, 8] for a detailed exposition of truth-value semantics.

There remains to say something philosophical about the truth-values in truth-value semantics. This type of semantics coheres with different views of truth:

(1) Truth-values are primitive, undefinable logical objects.

(2) Truth and falsehood are theoretical terms of the metalanguage, (weakly) implicitly defined by the semantic rules for logically complex sentences.

(3) Truth and falsehood are theoretical terms of the metalanguage. They are (weakly) implicitly defined by the semantic rules for logically complex sentences and partially interpreted by our attitudes of firm belief and firm disbelief with regard some sentences of the object language.

If the attitudes of firm belief and firm disbelief are classically inconsistent, classical logic is not (descriptively) applicable. This is is not a disaster. At least two options remain: (i) We have to emphasize the normative status of classical logic, and ask the speakers to make their beliefs consistent. (ii) We go non-classical and/ or do belief revision theory in the style of Gärdenfors [5]. Degrees of belief may then be seen as a generalization of the two semantic values for sentences in classical logic. Of course, more needs to be said about the philosophical interpretation of truth-values, but the above hints may suffice for now.

Why could it be promising to take truth-value semantics seriously? Obviously, if we use truth-value semantics right away for the object language, then we get rid of the problem of unintended interpretations of a theory T that cannot be ruled out as unintended. (This problem is at the core of Putnam's paradox.) Take Hans's toy example (p. 264):

- $\Sigma = \{c, d\}$
- $T = \{c \neq d, \forall x((x = c) \vee (x = d))\}$.

In the framework truth-value semantics, it is not even conceivable that Niels and Mette interpret T differently, provided they understand the language of T. The models of T are then simply certain truth-value assignments to the atoms of $L(\Sigma)$ and $L^{+}(\Sigma)$. Niels and Mette cannot diverge as to which truth-value assignments to the atoms count as models of T. To be more precise: there are no two models of T that are inconsistent with one another from the perspective of the metalanguage. In the Tarski semantics of T, by contrast, we have a model where c refers to a and one where c refers to b. (a and b are names of the metalanguage.) Hence, we can tentatively conclude that, in the framework of truth-value semantics, the language of T can be learned, mostly, within T. To be more precise: the truth-value semantics of T helps us learn what it means to firmly assert and to negate a sentence as regards the logical consequences of such an assertion or negation. But there is no need to further interpret or to learn the reference of the descriptive symbols c and d.

This tentative conclusion seems to align quite well with Hans' philosophical views concerning Putnam's paradox:

> Imagine two people, Mette and Niels, both of whom accept T, and both of whom think that the world is the set $\{a,b\}$. And yet, Mette says that c denotes a, whereas Niels says that c denotes b. Do Mette and Niels disagree? The answer is yes and no.
>
> *We have already misdescribed the situation.* Mette cannot say that 'c denotes a', because a is not a name in her language. Similarly, Niels cannot say that 'c denotes b'. (p. 264; emphasis added)

Obviously, in the framework of truth-value semantics, there is no temptation to misdescribe the situation in the first place. However, Hans also describes situations where a Tarski-style metalanguage allows to make some interesting and non-trivial distinctions:

> 1. The metalanguage describes the world in finer-grained language than the object language.
>
> 2. Distinctions that are not made by the object language are not significant for the kinds of explanations that the theory T gives. (p. 262)

Trying to be a liberal and pluralist Carnapian, I do not mean to suggest to dispense with Tarski semantics all together. In particular, it is not my intention to criticize the category-theoretic treatment of Tarskian model-theoretic semantics. For truth-value and Tarski semantics may well coexist. If the metalanguage allows us to describe the world in a more finer-grained way, we may well take advantage of this language. However, once we start looking for a semantics of the metalanguage (or a semantics of the metalanguage of the metalanguage), we should think about truth-value semantics. This move seems promising for two reasons:

(1) In response to Putnam, we can point out: it does not hold true anymore that 'we are in the same predicament with respect to the metalanguage that we are in with respect to the object language, ...' (Putnam [10, p. 482]).

(2) While there may still be a hierarchy of metalanguages in truth-value semantics (since we can always formalize the metalanguage and ask for a semantics), there is no infinite hierarchy of the interpretation of descriptive symbols any more.

So, let me conclude with a question to Hans: could truth-value semantics merit further consideration in the context of Putnam's paradox and related philosophical problems?

References

[1] Andreas, H. (2010). A Modal View of the Semantics of Theoretical Sentences. *Synthese* **174**(3): 367–383.

[2] Carnap, R. (1958). Beobachtungssprache und theoretische Sprache. *Dialectica* **12**: 236–248.

[3] Carnap, R. (1961). On the use of Hilbert's ϵ-operator in scientific theories. In *Essays on the foundations of mathematics*, edited by Y. B.-H. et al., Jerusalem: The Magnus Press. 156–164.

[4] Carnap, R. (1975). Observational Language and Theoretical Language. In *Rudolf Carnap. Logical Empiricist*, edited by J. Hintikka, Dordrecht: D. Reidel Publishing Company. 75–85.

[5] Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge, Mass.: MIT Press.

[6] Hintikka, J. (1962). *Knowledge and Belief: an Introduction to the Logic of the two Notions*. Ithaca: Cornell University Press.

[7] Leblanc, H. (1976). *Truth-Value Semantics*. Studies in Logic, Amsterdam: North-Holland Publishing Company.

[8] Leblanc, H. (1983). Alternatives to Standard First-Order Semantics. In *Handbook of Philosophical Logic: Volume I: Elements of Classical Logic*, edited by D. Gabbay and F. Guenthner, Dordrecht: Springer Netherlands. 189–274.

[9] Lewis, D. (1970). How to Define Theoretical Terms. *The Journal of Philosophy* **67**(13): 427–446.

[10] Putnam, H. (1980). Models and Reality. *Journal of Symbolic Logic* **45**(3): 464–482.

[11] Ramsey, F. P. (1931). Theories. In *The Foundations of Mathematics and Other Logical Essays*, edited by R. B. Braithwaite, London: Routledge and Kegan Paul. 212–236.

[12] van Fraassen, B. (1969). Presuppositions, Supervaluations and Free Logic. In *The Logical Way of Doing Things*, edited by K. Lambert, New Haven: Yale University Press. 67–92.

[13] Wright, C. (1993). *Realism, Meaning and Truth*. Oxford: Blackwell, 2nd edn.