ON THE "SEMANTICS" FOR LANGUAGES WITH THEIR OWN TRUTH PREDICATES

In *Truth, Definition and Circularity* (A. Chapuis and A. Gupta, eds.)
Indian Council of Philosophical Research, New Delhi, 2000, 217-246.

Philip Kremer, Department of Philosophy, McMaster University

In the 1920s, logical positivists were skeptical of the notion of truth. For one thing, the liar's paradox seemed to show that truth is inconsistent. Other sources of skepticism were, as Soames [26] notes, "the frequent use of truth in metaphysical discussions, the tendency to confuse truth with epistemological notions like certainty and confirmation, and the inability to see how acceptance of a truth predicate could be squared with the doctrine of physicalism and the unity of science."

On the other hand, the notion of truth seemed useful, even for scientific purposes. We would like to be able to assert that all the axioms of arithmetic are *true*, without asserting each of them. We might want to state general principles such as "valid arguments lead from true premises to true conclusions." What was wanted was a tool for "semantic ascent": As Etchemendy [2] puts it, the power to "recover or reassert a proposition [or a set of propositions] but to do so indirectly, without actually using a sentence that would, in the more usual fashion, express that same proposition."

In 1933, Tarski [27] provided a general method for giving a definition of truth for formalized languages. His method was widely regarded as a rehabilitation of truth: he had shown how to give a mathematical definition of a notion of truth that was suitable for scientific purposes.

Famously, the liar's paradox prevented Tarski from extending his methods to languages that have their own truth predicates, or that express their own truth-concepts. Both Tarski's original definition of truth, and its successor, the definition of truth-in-a-model, were not carried out in the language under investigation: *truth* is defined *in* a metalanguage for sentences *of* an object language. Only in the mid 1970s do we see papers—notably, Martin and Woodruff [21] and Kripke [19]—that extend Tarski's methods to languages that do, in some sense, contain their own truth predicates, even languages that display liar-like phenomena. These papers generated an industry aimed at providing "semantics" for languages expressing their own truth-concepts.

In the present paper, I want to ask a question that has not often been asked: In what sense are the so-called "semantics" offered by this newer work genuinely a "semantics"? This question can be asked of the work of Tarski and his peers as well, and various answers have been given. Part of the problem is, of course, with the notion of a "semantics". I do not want to hang too much on that word. Rather, I want to ask: In what sense does the work on the Liar's Paradox extend the accomplishment of Tarski's and his peers' work on non-paradoxical languages?

In the present paper, I concentrate on the fixed-point semantics introduced by Martin and Woodruff and by Kripke. By way of introduction, in §1, I begin with a discussion of the Tarskian and neo-Tarskian definitions of truth. In §2, I move on to the fixed-point semantics. In §3, I raise a central interpretive difficulty for this semantics.[1] In §4, I consider the semantics of vagueness, which seems beset by the same difficulty. In §5, I consider an objection to the effect that I have overlooked a natural interpretation of the fixed point semantics. §6 is by far the most speculative section of the paper, and I do not want to hang too much on the details: I consider two interpretations of Tarski, to see whether these can be applied to the fixed point semantics. In §7, I conclude.

§1. Tarskian and neo-Tarskian definitions of truth. Contemporary logic textbooks often introduce the notion of truth in three steps. They first introduce uninterpreted first order object languages as purely syntactic objects, each language specified by a basic vocabulary of names and predicates and such: this vocabulary generates a set of well-formed terms and well-formed formulas. We are then introduced to a notion of a *model*, *interpretation* or *structure*: if S is an uninterpreted language then a *model* for S is an ordered pair M = ⟨D, I⟩ where D is a non-empty domain of discourse and I is a function assigning an object in D to each name of S, a subset of D to each one-place predicate, and so on. Given a name or predicate X, we often think of I(X) as the *referent* of X. Often, the ordered triple L = ⟨S, D, I⟩ is thought of as an *interpreted*

[1] I am not, by the way, going to complain that the metalanguage is more expressive than the object language or that we are not shown how to give a semantics for a universal language, capable of expressing its own semantic notions. Nor am I going to raise the related issue of the strengthened liar.

language. We are then given the classic recursive definition of a sentence of an interpreted language L = ⟨S, D, I⟩ being *true* in L (or, equivalently, of a sentence in an uninterpreted language S being *true in a model* M = ⟨D, I⟩). This is done by giving a purely mathematical definition of a set $TRUE_L$ of sentences of L. We are not typically presented with any criteria of adequacy for such a definition, but we can draw an intuitive one out of Tarski's work.

Tarski [27] does not give a definition of truth in a model.[2] Tarski gives a definition of truth for a first order language which he assumes either to be a fragment of the metalanguage or to have a preordained translation into the metalanguage. Tarski's metalanguage is English (or Polish) with the vocabulary of standard set theory. Fixing the object language, Tarski gives a mathematical definition of a set, TRUE, of object-language sentences. Tarski's definition differs from the standard neo-Tarskian definition in notable ways: (1) each object-language constant is implicitly assumed already to have a referent, and there is no possibility envisioned of re-interpreting a constant by giving it a new referent;[3] and (2) there is no presupposition that the object-language quantifiers range over a *set* rather than a *class*, whereas the definition of a *model* requires the domain of discourse to be a set.[4]

---

[2]According to Hodges [16], the first time the notion appears in print in Tarski's work is in 1952 in [31], and the first time Tarski explicitly defines it in print is in 1957 in [33]. Hodges provides an excellent discussion of the differences between Tarski's notion and the contemporary notion, and of the relationship between them.

[3]In the contemporary scenario, if we are thinking of the definition of truth as a definition of truth for interpreted languages (rather than a definition of truth-in-a-model for uninterpreted languages) then the referents of the constants are also built into the language.

[4]The Tarskian set-up also allows predicate symbols to have, in effect, classes as referents. The reason is that the base clauses of the definition of truth do not "pass through" a function assigning an object to each name and a set to each predicate. Rather the base clauses are given directly. For example, suppose that the object language has a binary relational symbol R (meaning "less than") and two names, o (zero) and i (one). Then the base clause for the truth of Roi is not something like "Roi ∈ TRUE iff ⟨I(o), I(i)⟩ ∈ I(R)", but rather "Roi ∈ TRUE iff 0 < 1". Tarski's great innovation was to define truth for sentences in a language with quantifiers by first defining when a sequence of objects ⟨z_1, z_2, ...⟩ satisfies a formula α. If the object-language variables are x_1, x_2, ..., the base clauses of the recursive definition of satisfaction, for formulas Rx_ix_j will be "⟨z_1, z_2, ...⟩ satisfies Rx_ix_j iff $z_i < z_j$". Note that nothing in this definition requires the "referent" of R to be a set, since there is not talk of the referent of R. It is something of an anachronism to think of Tarski as having defined truth in terms of reference of constants and n-ary predicates.

Tarski's definition divides the set of all object-language sentences into two mutually exclusive subsets, TRUE and FALSE. Presented with any two-way partitioning of the sentences of an object language, one might wonder why *this particular* partitioning is interesting. Tarski insists that any adequate definition of truth must satisfy *Convention (T)*. That is, all the famous T-biconditionals must be provable:

(T) $x \in$ TRUE if and only if p,

where x is replaced by a name or structural description of an object-language sentence, and p is replaced by that sentence itself if the object language is a fragment of the metalanguage, and by a translation of that sentence into the metalanguage otherwise.[5] The reason why Tarski's particular partitioning of the sentences is interesting is that Convention (T) is, in fact, satisfied (but see footnote 5).

The fact that the T-biconditionals follow from Tarski's definition allows him to say that he has, in some sense, given a mathematical definition of truth. Whatever else, his definition does seem to allow the semantic ascent that logical positivists wanted. So the precise mathematical notion "$x \in$ TRUE" can do at least some of the work we wanted done by a notion of truth. One of the questions I will ask about the semantics for languages with their own truth predicates is this: can the metalinguistic notions introduced in that semantics be put to the same use as Tarski's notion of truth?

Contemporary work on the liar's paradox deals with object languages much more like the abstract interpreted languages set up at the beginning of this section, than with object languages as conceived by Tarski in [27]. We can articulate (strictly speaking, in the metametalanguage) a version of Convention (T) suitable for abstract interpreted languages. Fix an interpreted object

---

[5]Tarski notes that, strictly speaking, this criterion must be stated in the metametalanguage. Furthermore any proof that a particular definition of TRUE meets this criterion must be carried out in the metametalanguage. (I owe this observation to Peter Woodruff.) If we restrict ourselves to the metalanguage, the best we can do is show particular T-biconditionals for particular sentences.

language $L = \langle S_L, D_L, I_L \rangle$.[6] First translate every term constant c of the object language to the term 'I_L(c)' of the metalanguage, and every predicate constant G of the object language to the term 'I_L(G)' of the metalanguage. Then translate every atomic object language sentence Gc to the metalanguage sentence '$I_L(c) \in I_L(G)$'.[7] For each object-language sentence α, this produces a metalanguage sentence translation(α). We then say that a definition of $TRUE_L$ satisfies Convention (T_L) iff for every sentence α of the object language L, every sentence of the following form can be proved, in the metalanguage:

(T_L) $\alpha \in TRUE_L$ iff translation(α).

We will call these the T_L-biconditionals. Again, for any fixed object language L, Convention (T_L) is satisfied.[8]

Before moving on to the fixed point semantics, I want to note that the standard neo-Tarskian definition of the set $TRUE_L$ depends on no semantic or metaphysical notions: the definition uses a little syntax, a little set theory and concepts expressible in L. It might seem that we also need the notion of *reference*, since the base clauses of the recursive definition of $TRUE_L$ look something like this:

(1) $Gc \in TRUE_L$ in L iff $I_L(c) \in I_L(G)$,

the right side of which might be taken as a formalization of

(2) The referent (in L) of c is in the set which is the referent (in L) of G.

But, as Soames [26] points out, I_L is a purely mathematical object, so that even the base clauses of the definition of $TRUE_L$ do not rely on any undefined semantic notions. Thus we get a

---

[6]Strictly speaking, we will articulate a different Convention (T_L) for each interpreted object language L.

[7]Strictly speaking, this procedure assumes that the metalanguage has a name for every constant in the object language—or at least that each constant in the object language can be uniquely picked out in the metalanguage, e.g. by being the unique object satisfying such and such conditions expressible in the metalanguage. So our new version of Convention (T) can only be given subject to this restriction.

[8]I am not sure whether this can be made mathematically precise. But as an intuitive criterion, I think that it is clear enough, and it is clear enough that the criterion is satisfied.

definition of truth suited to the logical positivists' purposes and according to logical positivists' scruples.

§2. Fixed point semantics. As mentioned, the neo-Tarskian definition of truth-in-L is not given for languages that, in some sense, are able to express their own truth-concepts. Suppose that S is an uninterpreted language with a distinguished predicate T. And suppose that we want to *interpret* S by defining an interpreted language L = ⟨S, D, I⟩ in such a way that T expresses truth-in-L. Suppose also that for every sentence α of S there is a quote-name ‘α' in S.

Example 1. Let S be an enriched version of the language of arithmetic, with variables, quantifiers, and connectives; two constants, o and λ; three function symbols, s, + and ×; the identity sign, =; a unary predicate, Num; and a unary predicate, T. S also has quote names: for each sentence α, ‘α' is a closed term. Here are some terms of S: o, λ, sso, (o + λ), ‘Num(o × λ)', ‘¬Tλ'.

Before trying to interpret S so that T expresses truth, we might start with an interpretation of the T-free fragment of S. We might then see whether there are any ways to extend this supposedly unproblematic fragment to an interpretation of the whole language S, in such a way that T means "true-in-L" in the resulting interpreted language L. We will take a *ground model* to be a classical interpretation of all the names and predicates of S, *except the special predicate* T. We will also insist both that the domain of discourse of a ground model contain all the sentence of S and that the object assigned to the quote-name ‘α' always be the sentence α.

Example 2. Beginning the with uninterpreted language of Example 1, let the ground model be M = ⟨D, I⟩ with D = $\mathbb{N}$ ∪ {α: α is a sentence of S}; I(o) = 0; I(Num) = $\mathbb{N}$; I(λ) = ¬Tλ; I(‘Num(o × λ)') = Num(o × λ), etc.

We would like to extend the ground model to a model for the whole language to get an interpreted language L in which the predicate T means true-in-L. Suppose we insist that the ground model be extended to a classical model. And suppose that we want to give the standard

neo-Tarskian definitions of the sets $TRUE_L$ and $FALSE_L$ for the resulting interpreted language L. If T is to mean true-in-L, then a minimal condition on the resulting sets should be as follows:

(3) For each sentence α of L, T‘α' $\in$ $TRUE_L$ iff $\alpha \in TRUE_L$

(This implies that T‘α' $\in$ $FALSE_L$ iff $\alpha \in FALSE_L$.)

(This also implies that the extension of T is the set $TRUE_L$.)

Because of the liar's paradox, not every ground model is extendable in this way. Consider Examples 1 and 2. Assume, for a *reductio*, that M has been extended to a model $M' = \langle D, I' \rangle$ so that $L = \langle S, D, I' \rangle$ satisfies condition (3). Note that $I'(\lambda) = \neg T\lambda$. So, by condition (3), T‘$\neg T\lambda$' $\in$ $TRUE_L$ iff $\neg T\lambda \in TRUE_L$. Also $T\lambda \in TRUE_L$ iff T‘$\neg T\lambda$' $\in$ $TRUE_L$ since $I'(\lambda) = I'$(‘$\neg T\lambda$'). So $T\lambda \in TRUE_L$ iff $\neg T\lambda \in TRUE_L$. So $T\lambda \in TRUE_L$ iff $T\lambda \notin TRUE_L$. This is just a formalization of the liar's paradox, with $\neg T\lambda$ as the offending liar's sentence.

In giving a semantics for such a language, it is natural to put the liar's sentence $\neg T\lambda$ into neither the set $TRUE_L$ nor the set $FALSE_L$. Instead of classical interpreted languages, Martin, Woodruff and Kripke consider interpreted languages with truth-value gaps, languages whose sentences will be partitioned into three sets, $TRUE_L$, $FALSE_L$, and $NEITHER_L$.

In the general case, a gappy model for an uninterpreted language assigns to every name an object in the domain of discourse, and to every predicate two non-overlapping subsets of the domain of discourse, both an extension and an antiextension. We will say that a predicate has been given a *classical* interpretation if its antiextension is the complement of its extension, and that a gappy model is *classical* if all the predicates have been given a classical interpretation. Given a gappy interpreted language $L = \langle S, D, I \rangle$, we can define three sets of sentences $TRUE_L$, $FALSE_L$ and $NEITHER_L$ as follows (I will ignore the technology that Tarski introduces to deal with quantifiers):

Base clauses

$Gc \in TRUE_L$ if $I(c) \in$ the extension of G

$Gc \in FALSE_L$ if $I(c) \in$ the antiextension of G

$Gc \in NEITHER_L$ otherwise.

Recursive clauses for $\neg$ $\quad \neg\alpha \in \text{TRUE}_L$ if $\alpha \in \text{FALSE}_L$

$\neg\alpha \in \text{FALSE}_L$ if $\alpha \in \text{TRUE}_L$

$\neg\alpha \in \text{NEITHER}_L$ otherwise

Recursive clauses for $\vee$ $\quad (\alpha \vee \beta) \in \text{TRUE}_L$ if $\alpha \in \text{TRUE}_L$ or $\beta \in \text{TRUE}_L$

$(\alpha \vee \beta) \in \text{FALSE}_L$ if $\alpha \in \text{FALSE}_L$ and $\beta \in \text{FALSE}_L$

$(\alpha \vee \beta) \in \text{NEITHER}_L$ otherwise

There are other schemes for assigning "truth values" (including "Neither") to compound sentences. Above we used the Strong Kleene Scheme, which differs from the Weak Kleene Scheme in its interpretation of $\vee$ as follows:

Strong Kleene Scheme

$\vee$	T	F	N
T	T	T	T
F	T	F	N
N	T	N	N

Weak Kleene Scheme

$\vee$	T	F	N
T	T	T	N
F	T	F	N
N	N	N	N

A third scheme is van Fraassen's supervaluation scheme ([34]). Given a gappy interpreted language $L = \langle S, D, I\rangle$, say that a classical interpreted language $L' = \langle S, D, I'\rangle$ is a *precisification* of L iff the extension of each predicate in L is a subset of its extension in L', and the antiextension of each predicate in L is a subset of its antiextension in L'. Note that for a precisification L', the sets $\text{TRUE}_{L'}$ and $\text{FALSE}_{L'}$ can be defined using the classical definition, since L' is classical. We then define $\text{TRUE}_L = \{\alpha: \alpha \in \text{TRUE}_{L'}$ for all precisifications L' of $L\}$, $\text{FALSE}_L = \{\alpha: \alpha \in \text{FALSE}_{L'}$ for all precisifications L' of $L\}$, and $\text{NEITHER}_L = \{\alpha: \alpha$ is a sentence of S and $\alpha \notin \text{TRUE}_L$ and $\alpha \notin \text{FALSE}_L\}$. Note that even when $\alpha \in \text{NEITHER}_L$, we have $(\alpha \vee \neg\alpha) \in \text{TRUE}_L$. In the weak and strong Kleene schemes, if $\alpha \in \text{NEITHER}_L$ then $(\alpha \vee \neg\alpha) \in \text{NEITHER}_L$.

Suppose that we start with a language whose T-free fragment is given a classical ground model. Can we extend the ground model to a gappy model for the whole language, to get an interpreted language L in which the predicate T means true-in-L? A minimal condition on the sets TRUE_L, FALSE_L and NEITHER_L should be as follows:

(4) For each sentence α of the object language

$T\ulcorner\alpha\urcorner \in TRUE_L$ iff $\alpha \in TRUE_L$, and

$T\ulcorner\alpha\urcorner \in FALSE_L$ iff $\alpha \in FALSE_L$, and

$T\ulcorner\alpha\urcorner \in NEITHER_L$ iff $\alpha \in NEITHER_L$.

This is equivalent to

(5) $TRUE_L$ = the extension of T, and

$FALSE_L$ = the antiextension of T.

If an interpreted language satisfies these conditions, we say that it is a *fixed point*.[9]

Martin and Woodruff's and Kripke's main result is that every ground model can be extended to a fixed point.[10] In fact, most ground models can be extended to a multitude of fixed points. Thus, the fixed point approach yields a number of plausible interpretations of T, and generates the hope that we have the beginnings of a semantics for languages that can express their own truth-concepts.

Before moving on to our interpretive difficulty, I want to note that the literature contains two main proposals for interpreting the fixed point semantics. On the most common proposal, if you begin with a language whose T-free fragment is interpreted with the ground model M, then the correct interpretation of T is given by the *least fixed point* associated with this or that evaluation scheme, where the fixed points are ordered in a particular way. (See [15], [12], [1], [20], [24], [17], and [25].) M. Kremer [18] sees this proposal as motivated by the "supervenience of

---

[9]Suppose that we fix an uninterpreted object language S, with quote names for the sentences, and a distinguished predicate T. Also suppose that we fix a ground model $M = \langle D, I\rangle$ for S. Consider an interpreted language $L = \langle S, D, I'\rangle$ where I′ extends I by giving a gappy interpretation to T. Define $J(L) = \langle S, D, J(I')\rangle$ by letting J(I′) agree with I and I′ on all constants other than T, and by letting J(I′) assign the following extension and antiextension to T: extension = $\{\alpha: \alpha \in TRUE_L\}$; antiextension = $\{\alpha: \alpha \in FALSE_L\}$. Then the fixed points are precisely the fixed points of the "jump" operator J: they are the gappy interpreted languages L such that J(L) = L.

[10]This result holds on the assumption that the scheme used to assign a truth-value to compound sentences has a property called "monotonicity". The monotonicity of the scheme ensures the monotonicity (in a certain sense) of the jump operator (see footnote 9), and this ensures the existence of fixed points. The weak and strong Kleene schemes and the supervaluation scheme are monotonic, as are some variants on the supervaluation scheme. Gupta and Belnap [14] provide a nonmonotonic scheme for which the fixed point theorem holds.

semantics", an intuition according to "once we are given the interpretation of all the non-semantical constants, a further interpretation of the truth predicate is redundant". On this proposal, if T is to mean truth, then its interpretation should be completely determined by the ground model. The supervenience of semantics does not by itself dictate the choice of the least fixed point as the correct interpretation of T: it merely dictates that some particular fixed point be assigned to each ground model. The least fixed point has often simply seemed the most natural.

Kremer decisively argues that Kripke [19] does not endorse this proposal. He furthermore argues the supervenience proposal is in tension with another redundancy intuition, the one Kremer sees as motivating the "fixed point conception of truth". According to this intuition, the intuitive concept of truth is completely captured by Kripke's formulation in [19]: "we are entitled to assert (or deny) of a sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself." This conception of truth favours no particular fixed point, since this formulation holds equally well of all fixed points.

One way to see the difference between the two proposals is to consider two communities C and C′, speaking syntactically identical interpreted first order languages L and L′. Suppose furthermore that both languages contain a predicate T, and that the T-free fragments of L and L′ are represented by the same ground model M. Finally suppose that L and L′ are distinct fixed points, relative to M and to an appropriately chosen evaluation scheme. On the nonsupervenience proposal, both communities C and C′ are using the predicate T to express truth: C is using T to express truth-in-L and C′ is using T to express truth-in-L′. On the supervenience proposal, either C is using T to express a concept other than truth-in-L or C′ is using T to express a concept other than truth-in-L′: since both L and L′ are fixed points sharing a ground model, in at most one of them can T mean truth.

Without considering the relative merits of these two proposals, we move on to our interpretive difficulty.

§3. The difficulty. Suppose we are presented with a fixed point: an interpreted language L, where every predicate except T is given a classical interpretation and T is given a gappy interpretation. Furthermore we have condition (4), above. Since the point I want to make in this section is independent of the supervenience of semantics, we will not worry whether L is, for example, the least fixed point relative to the ground model and appropriately chosen evaluation scheme.

In this section, my main contention will be

(6) the metalinguistic expression '$x \in TRUE_L$' does not express the concept "x is a true sentence of L" (and does not even express an extensionally equivalent concept).[11]

The metalanguage is the language we are giving our semantics *in*. Since Tarski, the main tradition in semantics has been to define, in the metalanguage, a notion of truth for the object language. In Tarski's setting, we can think of ourselves as having done just this, because of the T-biconditionals. If contention (6) is right, then given a fixed point L—a language that purportedly expresses its own truth-concept—we do not, despite appearances, have in the metalanguage, a notion of truth for the object language.

This does not mean that L fails to express its own truth-concept. Rather it means that we cannot express L's truth-concept in the metalanguage with the formula '$x \in TRUE_L$'. Now, it is not written in stone that a semantics must proceed by defining, in the metalanguage, a notion of truth for the object language. Perhaps we can assign abstract objects to the expressions of the object language—or partition the object-language sentences into two, three or more subsets—in a way that is still illuminating. Tarski has an answer as to why his particular two-way partitioning of the sentences is useful: one of the parts is the set $TRUE_L$ for which we can prove the T-biconditionals. We are owed an answer as to why this new three-way partitioning of

[11]Henceforth, we will ignore the difference between expressing a concept C and expressing a concept extensionally equivalent to C.

sentences is interesting—especially if contention (6) is right that '$x \in TRUE_L$' does not express, in the metalanguage, the concept of truth in the object language L.

Now to argue for contention (6). Recall that the general project we are considering is to provide a semantics for languages expressing their own truth-concepts. This project *demands* that the object-language formula Tx express the concept "x is true in L". So it suffices to argue that the metalanguage formula '$x \in TRUE_L$' does not express the same concept as the object-language formula Tx.

I will borrow from Gupta and Belnap [14] the pseudo-technical notion of the *signification* of an expression: "the *signification* of an expression in a world w is an abstract something that carries all the information about the expression's extensional relations in w." This is supposed to be a generalization of the classical notions of reference for singular terms, and extension for predicates. It seems that the *signification* of the object language predicate T—i.e. the abstract object to which this predicate is assigned—is simply the ordered pair consisting of T's extension and antiextension. The metalanguage predicate '$x \in TRUE_L$' seems to have the same *extension* as T. But it has a different antiextension: the liar sentence, $\neg T\lambda$, is in the antiextension of '$x \in TRUE_L$' but not in the antiextension of T. Thus the metalanguage formula '$x \in TRUE_L$' does not have the same signification as the object-language formula Tx, and so does not express the concept "x is true in L".

A different way to see the point is as follows. First note that we are presented with a three-way partitioning of the sentences of L, much as Tarski would have presented us with a two-way partitioning if L were classical. Given any partitioning of the sentences of L, one might wonder why it is interesting. As already mentioned, Tarski's two-way partitioning is interesting because we can prove the T-biconditionals (or the T_L-biconditionals), given in §1, above. But there is no similarly clear answer as to the significance of the current three-way partitioning.

We do not seem to have, for example, the T_L-biconditionals

(T_L) $\alpha \in TRUE_L$ iff translation(α).

To see this, consider the uninterpreted language S and the ground model M = ⟨D, I⟩ of Examples 1 and 2 in §2, above. Let L = ⟨S, D, I′⟩ be a fixed point, where I′ extends I. If '$x \in \mathrm{TRUE}_L$' and Tx both express "x is a true sentence of L", then we get a false T_L-biconditional by considering a particular example:

Object language sentence	$T\lambda$
English translation of $T\lambda$	$I(\lambda) \in \mathrm{TRUE}_L$.
Object language sentence	$\neg T\lambda$
English translation of $\neg T\lambda$	$I(\lambda) \notin \mathrm{TRUE}_L$.
T_L-biconditional for the sentence $\neg T\lambda$	$\neg T\lambda \in \mathrm{TRUE}_L$ iff $I(\lambda) \notin \mathrm{TRUE}_L$.
The T_L-biconditional is equivalent to	$\neg T\lambda \in \mathrm{TRUE}_L$ iff $\neg T\lambda \notin \mathrm{TRUE}_L$.

These two arguments—from the difference in the signification of '$x \in \mathrm{TRUE}_L$' and of Tx and from the failure of the T-biconditionals—not only suggest that TRUE_L does not express truth-in-L; they leave it unclear what the interest is of the three-way partitioning of the sentences of L into the subsets TRUE_L, FALSE_L and $\mathrm{NEITHER}_L$.

I want to consider three complaints against this line of argumentation:

Complaint 1. Although '$x \in \mathrm{TRUE}_L$' does not express truth-in-L, the three-way partitioning of the sentences is still interesting since we have condition (4), at the end of §2, as an analogue of Convention (T).[12]

Complaint 2. The point is well-taken, but it has nothing special to do with the liar's paradox: the problem can be raised in any semantics that uses truth-value gaps.

Complaint 3. I have interpreted the fixed point semantics by concentrating on the goings on in a *single* fixed point. But, on one interpretation, it is *the totality* of fixed points extending a ground model that is significant.

Consider Complaint 1. Condition (4) shows that the three-way partitioning of the sentences into TRUE_L, FALSE_L and $\mathrm{NEITHER}_L$ always puts α and T'α' in the same category. But this

[12]Note that McGee's [22] "material adequacy condition" is the analogue, in his context, of condition (4).

alone does not show that the three-way partitioning is interesting or illuminating. It is easy to come up with completely boring three-way partitions that always puts α and T‘α' in the same category.[13] Perhaps what is more relevant is that condition (4) is equivalent to condition (5):

(5) $TRUE_L$ = the extension of T and

$FALSE_L$ = the antiextension of T.

But taking condition (5) to be a suitable analogue of Convention (T) assumes that the notions of extension and antiextension are semantically significant. But, in a way, this is exactly what is at issue: the role of the notions of extension and antiextension is, after all, to generate the definitions of $TRUE_L$ and $FALSE_L$, and we are now questioning of the semantic significance of these defined notions.

I should add that condition (4) can be considered an analogue of Convention (T), once we have established that the metalinguistic notions, $TRUE_L$, $FALSE_L$, and $NEITHER_L$, are semantically significant. The satisfaction of condition (4) might then be a necessary condition for T to mean truth: we might require the semantics to treat α and T‘α' in the same way, in at least some respects. There is an important disanalogy with Convention (T): Convention (T) demands some kind of equivalence between an object-language sentence and a metalanguage sentence, whereas condition (4) requires some kind of equivalence between two object-language sentences.

I will consider Complaint 2 in §4 and Complaint 3 in §5.

§4. Other gappy semantics. There are other linguistic phenomena for which truth-value-gap semantics have seemed appropriate, notably the phenomenon of vagueness.[14] An oversimplified semantics for a first order object language with vague predicates (but precise names and quantifiers) might be given by considering the gappy models for uninterpreted languages

[13]Divide the sentences into three sets, X = {α: α is a sentence and the predicate G occurs in α} (perhaps in the scope of quotation marks), Y = {α: α is a sentence and the predicate H occurs in α} and Z = {α: α is a sentence and neither G nor H occurs in α}.

[14]There is also the phenomenon of nondenoting singular terms, which I will ignore for the purposes of this paper.

suggested in the middle of §2, above. Recall that in a gappy model, a predicate is assigned not only an extension but an antiextension, where these do not overlap. The idea is that with a vague predicate G, there are clear cut cases of objects that are G (these are in the extension of G), there are clear cut cases of non-G objects (these are in the antiextension) and there are borderline cases (these are in neither).[15] Given an interpreted vague first order language L, the sets $TRUE_L$, $FALSE_L$ and $NEITHER_L$ (or $INDETERMINATE_L$) can be defined by the weak or strong Kleene scheme, the supervaluation scheme, or by some other scheme.

Fine [11] and McGee and McLaughlin [23] prefer a variation on the supervaluation scheme. If we consider all precisifications of the language L, then, for any two vague predicates G and H, if G and H share some borderline cases then the sentences $\forall x(Gx \rightarrow Hx)$ and $\forall x(Hx \rightarrow Gx)$ come out indeterminate rather than true. But we might want 'all bachelors are unmarried' always to come out true, even if 'bachelor' and 'unmarried' are both vague and share borderline cases. One way around this is to add to the notion of an interpreted language a set Δ of constraints (what Fine calls "penumbral connections"): Δ might simply be a set of object-language sentences that are considered analytic. Say that a precisification that satisfies these constraints is *acceptable*. In particular, if Δ is a set of object-language sentences, then a precisification L′ is acceptable iff $\Delta \subseteq TRUE_{L'}$. $TRUE_L$ is redefined so that $\alpha \in TRUE_L$ iff $\alpha \in TRUE_{L'}$ for each acceptable precisification L′ of L. Similarly, $FALSE_L$ is redefined so that $\alpha \in FALSE_L$ iff $\alpha \in FALSE_{L'}$ for each acceptable precisification L′ of L. And $\alpha \in NEITHER_L$ iff $\alpha \notin TRUE_{L'}$ and $\alpha \notin FALSE_{L'}$.[16] Similarly for $FALSE_L$ and $NEITHER_L$. One reason McGee and McLaughlin

[15]One oversimplification is that we are not obviously allowing for borderline cases or borderlineness, cases of higher order vagueness.

[16]McGee and McLaughlin insist that "the idea that a sentence is definitely true if and only if it is true in every member of a certain class of models needs to be clearly distinguished form the thesis that a sentence of a vague language is definitely true if and only if it is true in all the fully precise languages that can be got by making the given language precise." They consider the second thesis insupportable: "we cannot get from English to a fully precise language merely by deciding what to say about some hard cases which English usage leaves unsettled" (p. 227). I have set everything up in terms of interpreted languages rather than uninterpreted languages and their

prefer the supervaluation scheme is that, as opposed to the Kleene schemes, it preserves classical logic: all classical theorems are in $TRUE_L$ for every interpreted language L.

The question arises whether '$x \in TRUE_L$' expresses the concept "x is a true sentence of L". We cannot apply considerations like those of §3, since these considerations depend on the existence of a truthlike predicate in the object language. McGee and McLaughlin do, however, maintain that there is an important sense in which '$x \in TRUE_L$' does not express the intuitive concept of truth (in L): they take it to express a related concept, *definite truth*.[17] They present interesting differences between the behaviour of '$x \in TRUE_L$' and 'x is a true sentence of L' (understood disquotationally). They argue, for example, that simple truth distributes over disjunction—$(\alpha \vee \beta)$ is true iff α is true or β is true—while definite truth does not. Similarly, they argue that 'Harry is bald' is either true or false, but they note that 'Harry is bald' is neither definitely true nor definitely false.

Granted that $TRUE_L$ captures a new notion of definite truth rather than truth, what is the significance of this new notion? McGee and McLaughlin suggest that it captures a notion of truth whose underlying principle is in conflict with the disquotation principle, as least when it comes to vague discourse: "the *correspondence principle* tells us that the truth conditions for a sentence are established by the thoughts and practices of the speakers of the language, and that a sentence is true only if the nonlinguistic facts determine that these conditions are met."

There are two problems I wish to raise for this semantics. Firstly, Tarski had a criterion of adequacy, Convention (T), for his notion of truth, a criterion intimately related to the disquotation principle. We are now trying to capture, with the formal definition of $TRUE_L$, a notion of *definite* truth that we have divorced from the disquotational principle. What criteria of adequacy

---

models. This might seem to favour the second thesis, but it is actually neutral between the two approaches. While the gappily interpreted language L might be seen as a formalization of a language in use, the precisifications of L can be seen as mathematical abstractions useful for giving a semantics for L.

[17]More precisely, they take it that the intuitive concept of truth bifurcates, upon philosophical scrutiny, into a disquotational concept of truth and a "semantic" concept of definite truth.

might we impose on a formal definition of definite truth? Whatever criteria we use, they will have to involve our intuitions about correspondence, which are much less succinctly expressible than is Convention (T). These criteria might involve the purposes we take the notion of definite truth to serve.

Secondly, Tarski's work was done in the context of logical positivism, which was highly skeptical about such notions as correspondence. Despite Tarski's claims to have captured the intuitive correspondence notion of truth, what really justified him was, as just mentioned, the fact that his defined notion satisfied Convention (T): Tarski's notion of truth allowed the semantic ascent that the logical positivists were after.[18] Can the notion of *definite* truth be justified without appeal to "correspondence" intuitions?

I will put off discussion of correspondence until §6, below. For now I point out that there might be a way to justify the notion of definite truth for vague languages, without appeal to intuitions surrounding correspondence, and on the basis of a notion of simple truth satisfying Convention (T). Note that if we use the supervaluation semantics, then definite truth in L is defined in terms of simple truth in all acceptable classical precisifications of L. And simple truth in these classical precisifications satisfies the appropriate version of Convention (T), because the precisifications are classical.[19] The notion of definite truth captures what it is that all the acceptable Convention-(T)-satisfying notions have in common. In some sense, definite truth is

---

[18]Various authors, notably Etchemendy [2], argue that this is all that Tarski was after. See §6, below.

[19]We might be reminded of McGee and McLaughlin's worries raised in footnote 16: they worry that thinking of precisifications as more precise languages is beset by a number of problems. In particular they propose that "the meanings of certain English terms requires that the terms be vague" (p. 229). They continue, "there is no plausible reason to believe that examining a class of languages whose rules contradict the rules of English will cast any useful light upon the semantics of English." I seem to be proposing precisely the analysis of "precisification" to which they object. But my point can be recast in terms that I think that they would find unobjectionable. They say that their position "does not require looking at a lot of different languages, but rather looking at a lot of different models ... of the vague language whose semantics we are trying to describe" (p. 228). If we think of precisifications as just that, then we have to recast our talk of "truth in an interpreted language L" in terms of "truth of a vaguely interpreted language L in a precise model M", and we would have to recast our statement of Convention (T_L), but this can be done.

the greatest common approximation of all the acceptable versions of Convention-(T)-satisfying truth.[20]

How does this tie into our discussion of the semantics for the liar's paradox? Recall Complaint 2 at the end of §3: that the worry I raised about the new metalinguistic notions has nothing special to do with the liar's paradox, since the problem can be raised in any semantic context in which truth-value gaps seem appropriate. In a sense this is right: the notion of definite truth ($TRUE_L$) for vague interpreted languages does not have the straightforward justification and utility that the notion of simple truth for precise interpreted languages has. But I claim that the analogous problem is more acute for the fixed point semantics of §2 than for the semantics for vague predicates. The reason is that there is no analogue for the fixed point semantics to the line of defence, sketched two paragraphs back, of the semantics for vague predicates. The problem is that in the case of the fixed point semantics, the precisifications of L do *not* in any clear way obey the appropriate versions of Convention (T). To see this, consider the uninterpreted language S and ground model $M = \langle D, I\rangle$ of Examples 1 and 2 in §2, above. Recall that $I(\lambda) = \neg T\lambda$. And let $L = \langle S, D, I_L\rangle$ be some fixed point relative to the supervaluation scheme. Finally, suppose that $L' = \langle S, D, I'\rangle$ is a precisification of L. The liar's paradox seems to provide a counterexample to the claim that L′ obeys the appropriate version of Convention (T), whether we take the object-language formula Tx to express, in L′, truth-in-L′ or truth-in-L, as the following argument shows.

Suppose the object-language formula Tx expresses, in L′, truth-in-L′. Then Convention (T) requires that for each sentence α of L′, $T\text{`}\alpha' \in TRUE_{L'}$ iff $\alpha \in TRUE_{L'}$. The argument from this to a contradiction has already been rehearsed: $T\text{`}\neg T\lambda' \in TRUE_{L'}$ iff $\neg T\lambda \in TRUE_{L'}$. Also $T\lambda \in TRUE_{L'}$ iff $T\text{`}\neg T\lambda' \in TRUE_{L'}$ since $I'(\lambda) = I'(\text{`}\neg T\lambda')$. So $T\lambda \in TRUE_{L'}$ iff $\neg T\lambda \in TRUE_{L'}$. So $T\lambda \in TRUE_{L'}$ iff $T\lambda \notin TRUE_{L'}$. On the other hand, suppose the object-language formula Tx expresses, in L′, truth-in-L. Then Convention (T) would require that for each sentence α of L′,

[20]This line of defence is unavailable for a semantics of vagueness that uses one of the Kleene schemes.

T‘α' ∈ $TRUE_{L'}$ iff α ∈ $TRUE_L$. We know that ¬Tλ ∉ $TRUE_L$. So, assuming Convention (T), T‘¬Tλ' ∉ $TRUE_{L'}$. So Tλ ∉ $TRUE_{L'}$. So ¬Tλ ∈ $TRUE_{L'}$ since L′ is classical. This last claim is simply not true of all precisifications L′ of L. Perhaps it is true of all *acceptable* precisifications (i.e. those obeying appropriately spelled out constraints). In that case, we would have ¬Tλ ∈ $TRUE_{L'}$ for all acceptable precisifications L′ of L. But in that case, ¬Tλ ∈ $TRUE_L$. But ¬Tλ ∉ $TRUE_L$.

So the line of defence for our semantics for vague predicates cannot be reused to defend the fixed point semantics, even when we are using the supervaluation scheme. So the unclarity surrounding the significance of the definition of $TRUE_L$ is even more worrisome in the case of the fixed point semantics than it is in the case of the semantics for vague predicates.

§5. One fixed point vs. the totality of fixed points. Suppose that a community speaks a language whose syntax is represented by the first order language S, which includes a quote name ‘α' for each sentence α, and a special one-place predicate T. On both the supervenience and the nonsupervenience proposals as I have considered them, the total language spoken by a community is represented by an interpretation of S, which gives a classical interpretation to the T-free fragment of S and a possibly gappy interpretation to T. Suppose that L is the resulting interpreted language. On the nonsupervenience proposal for interpreting the fixed point semantics, in order for T to mean truth-in-L, it suffices that L be a fixed point. On the supervenience proposal, L must be, for example, the least fixed point determined by the ground model interpreting the T-free fragment of S. On each of these proposals, to interpret T as truth is to assign it an extension and an antiextension determined by some particular fixed point.

But it might be noted that much of the interesting mathematical work on fixed points has centred on the *totality* of fixed points relative to a ground model, and on how they are structured and related to one another. Secondly, certain very intuitive definitions rely on the totality of fixed points. For example, let the truth-teller be the sentence that says of itself that it is true, just as the liar says of itself that it is false. Intuitively the truth teller is troublesome, but it is not

paradoxical: it does not seem to force inconsistency. Kripke defines a sentence to be *non-paradoxical* if it has a definite truth-value relative to *some* fixed point. The truth-teller turns out to be non-paradoxical while the liar is paradoxical. And this notion of non-paradoxicality requires us to consider all the fixed points.

The nonsupervenience proposal takes the totality of fixed points more seriously than the supervenience proposal. But a community's language is, even on the nonsupervenience proposal, represented by a single fixed point. This seems at odds with the idea that important semantic notions such as non-paradoxicality should be defined by quantifying over *all* fixed points: i.e. all languages, spoken or not, that both share the community's ground model and satisfy condition (4), above.

A third proposal, which is a kind of supervenience proposal, takes a community's language to be determined by the interpretation of its T-free fragment, and interprets T via a very abstract object: the class of all fixed points.[21] On this account, the semantics does not assign to each sentence a single value TRUE_L, FALSE_L or NEITHER_L. It seems each sentence is assigned a function from fixed points to these values. On this picture, no concept defined in the metalanguage even *purports* to be a metalinguistic characterization of truth.

But then the old question arises again: in what sense is this abstract assignment of objects to expressions to be counted as a semantics? Of what interest are the central semantic notions?

One answer is that it provides us with a theory of *good inference*. In fact, this answer might also be given by a non-supervenience theorist, since both she and the new totality-of-fixed-points theorist might quantify over all fixed points in giving a semantic account of good inference. This answer is encouraged by the logic of truth given by M. Kremer [18]. Relying on *all* the fixed points, Kremer gives a very natural definition of when one formula *logically implies* another: $\alpha \models \beta$ iff for each ground model M and each fixed point L extending the ground model, if $\alpha \in \text{TRUE}_L$ then $\beta \in \text{TRUE}_L$ and if $\beta \in \text{FALSE}_L$ then $\alpha \in \text{FALSE}_L$. Kremer shows that this relation

[21]Gupta and Belnap mention this proposal in footnote 40 in Chapter 3 of [14].

admits of a very natural sound and complete axiomatization. The proof that the semantics is significant is in the elegance of the logic that the semantics motivates.

On the other hand, Kremer's definition of logical consequence relies on the notion of the formula α being true relative to a fixed point. But we have already seen that the notion of truth-relative-to-a-fixed-point is not clearly a notion of *truth* at all. If logical consequence is defined as preservation of *this*—this thing which is not really truth—then it is unclear why we should be interested in logical consequence.

One might respond to this worry by claiming that we *already know* which logical inferences are the good ones, and that the semantics is merely a formal device for encoding them. This is the kind of story you might tell about modal logic: if we already know that the principles of S4 encode the right modal principles, we might want a mathematical tool for studying these principles. And the possible world semantics does the trick. Unfortunately, in the case of the logic of truth, we do not "already know" which inferences are good ones.[22] We are pretty sure that you can infer 'α is true' from α and vice versa. But our intuitions are unclear precisely where the semantics you choose makes a difference. If you choose the fixed point semantics with the strong Kleene scheme, for example, then (T‘α' $\vee$ $\neg$T‘α') is not a theorem. But if you choose the supervaluation scheme or the semantics of Gupta and Belnap [14], then (T‘α' $\vee$ $\neg$T‘α') is a theorem. The semantic notions seem to be doing more work than merely encoding already well-accepted inferential patterns.

§6. Two interpretations of Tarski. Tarski defined a notion of truth for certain formal languages, and his methods evolved into those used to define truth-in-L for classical interpreted first order languages L.[23] The importance of these notions lies in the T-biconditionals. Fixed point theorists try to extend the work of Tarski and his peers to formal languages that can express

[22]We probably do not know this in the modal case either.

[23]In my informal remarks, I am not going to be too careful to distinguish Tarski's notion of truth from its successor, the notion of truth-in-an-interpreted-language or truth-in-a-model.

their own truth-concepts. While we can use fixed points to define, in the metalanguage, a number of notions for the object language, we have been arguing that we do *not* define, in the metalanguage, a notion of truth for the object language. The question arises: What good are the notions that we *do* define? One answer was that they give a theory of good inference. This answer was found to be unsatisfactory.

We should recognize at least the mathematical similarity between the new notions and classical neo-Tarskian notion of truth-in-L. So it might help to consider philosophical interpretations of Tarski and the classical notion, in order to consider the philosophical significance of the work on fixed points.

Two principal interpretations of Tarski's work are interpretations of him as a kind of *deflationist* about truth (see §6.1); and interpretations of him as a *correspondence* theorist about truth (see §6.2). Etchemendy [2] proposes the first interpretation, and hints at the second. He takes Tarski to have replaced the intuitive notion of truth with a quasi-deflationist notion of truth, useful for some purposes. But he also suggests that Tarski's definitions can be reinterpreted as substantial claims about the ordinary intuitive notion of truth, in such a way that truth depends in part on the linguistic behaviour of speakers.

§6.1. Tarski the deflationist. One way to get a handle on Etchemendy's interpretation is to consider the following two biconditionals, where we are assuming that L is an appropriate fragment of English:

(7) ‘Snow is white' $\in$ $\mathrm{TRUE}_{\mathrm{L}}$ iff snow is white.

(8) ‘Snow is white' is a true English sentence iff snow is white.

(7) is theorem of set theory (extended with some syntax). Meanwhile, Etchemendy takes it that (8) expresses a substantive and contingent fact, a truth that depends on how speakers of English use their words. As Etchemendy insists, ‘Snow is white' might have been false even though snow were white—for example if ‘snow' referred to grass. The suggestion is that genuinely *semantic* notions are notions that depend in some way or other on the *use* people make

of words. Etchemendy then suggests that Tarski's recursive definitions of the set $TRUE_L$ can be converted into substantial semantic claims about the ordinary concept of truth, simply by replacing the occurrence of the phrase ‘x ∈ $TRUE_L$' with the phrase ‘x is true (in L)'.

There might be another way to make the point. If we are thinking of English as a certain *interpreted* language—in our technical sense of interpreted language[24]—then even (8) is a theorem of set theory. But Etchemendy's point can be captured by comparing a different pair of biconditionals:

(9) ‘Snow is white' ∈ $TRUE_L$ iff snow is white.

(10) ‘Snow is white' is true-relative-to-a-particular-speaker-P iff snow is white.

It is a contingent matter where (10) holds, but not because it is a contingent matter whether the word ‘snow' refers, in English, to snow. Rather, it is a contingent matter whether P is speaking a language in which the word ‘snow' refers to snow. On this line, if P uses the word ‘snow' to refer to grass, then P is not speaking English: the appropriate interpretation of the sentences that P utters does not assign snow to ‘snow'. This suggests a definition of truth-relative-to-a-speaker:

(11) X is $true_P$ iff X ∈ $TRUE_L$, where L is the language that P speaks.

This definition hinges on the idea of P speaking a particular interpreted language, an idea we will worry about later.

According to Etchemendy, Tarski is not attempting to capture the intuitive notion of truth—whether this is thought of as truth relative to a speaker or truth plain and simple. Rather Tarski is providing a replacement concept, which can be used for the purposes that the logical positivists wanted: for semantic ascent. It is worth noting that, for these purposes, we actually want a notion for which the truth of the T-biconditional is *not* contingent. To use an example from Field [8], when we say, "If the axioms of Euclidean geometry were true then the laws of physics would be different" we need a notion of truth that does not depend on how speakers use words in other possible worlds. Unfortunately, this deflationist interpretation of $TRUE_L$ is

[24]This, of course, is an idealization.

simply unavailable in the context of the fixed point semantics. Without Convention (T), $TRUE_L$ is simply not useful for semantic ascent.

§6.2 Correspondence. Tarski often insists that his notion of truth is a *correspondence* notion. (See [29] and [30].) This seems at odds with Etchemendy's deflationist reading of Tarski's work, a reading which is not uncommon.

Consider an abstract interpreted language $L = \langle S_L, D_L, I_L \rangle$. We could recast the standard neo-Tarskian definition of truth-in-L first by defining reference-in-L of terms and predicates, and then

by defining truth-in-L in terms of reference-in-L. The definition of reference-in-L would go as follows, for any term or predicate constant X:

(12) The referent_L of $X = I_L(X)$.

If L is a fragment of English, then (12) might be a handy way of summarizing a rather long list,

(13) The referent_L of ‘France' is France

The referent_L of ‘England' is England

The referent_L of ‘is a country' is the set of countries

...

since I_L is simply the set of ordered pairs {⟨‘France', France⟩, ⟨‘England', England⟩, ⟨‘is a country', {x: x is a country}⟩, ...}. We can then give the standard recursive neo-Tarskian definition of truth-in-L, starting with the base clauses,

(14) Gc ∈ $TRUE_L$ iff the referent_L of c ∈ the referent_L of G.

Does this definition provided a correspondence theory of truth-in-L? In some sense, yes: the notion of truth-in-L is defined in terms of correspondences between expressions of the language and objects in or subsets of the domain of discourse. But in that case, the conception of *reference* given in (12) or (13) is likewise a correspondence conception of reference. If this is all we mean by a "correspondence conception" of a notion like truth or reference, then there is

no conflict between correspondence conceptions and the quasi-deflationary conception attributed to Tarski by Etchemendy.

But something more robust is often expected from a correspondence conception of truth. We have already seen this, for example, from McGee and McLaughlin [23]: "the *correspondence principle* tells us that the truth conditions for a sentence are established by the thoughts and practices of the speakers of the language, and that a sentence is true only if the nonlinguistic facts determine that these conditions are met." Note that the neo-Tarskian conception of truth-in-L for interpreted languages L cannot accommodate the correspondence principle thus articulated. An interpreted language L is a mathematical object. If we fix an interpreted language L, then the truth-in-L of a particular sentence of L has no more to do with the thoughts and practices of any speakers than does the primeness of a particular natural number.[25]

Robust correspondence principles only make sense when we move from *abstract* semantics for languages conceived as mathematical objects, to semantics *applied* to real speakers. Suppose that P is a speaker and you want to define the set TRUE_P of sentences that are true-for-P. If P is fixed, then one strategy is to give a two-step definition: step 1 is to stipulate that P speaks such-and-such language L, and step 2 is to define $\text{TRUE}_P =_{df} \text{TRUE}_L$. Such a definition might be correct, or "materially adequate" to use Tarski's phrase—whatever the adequacy conditions are for a definition of "true-for-P". But there is a sense, derived from the classic discussion of Field [3], in which such a definition can be seen as problematic, if we want to know *what it is* for a sentence to be true-for-P.

Consider an analogy: you want to explain the notion of a "constitutional monarchy" to me. You define a set CM = {The UK, Sweden, The Netherlands, ...}, and then you define x to be a constitutional monarchy iff $x \in CM$. Your definition might be extensionally correct, but you have not successfully characterized the concept of a constitutional monarchy. Though I can tell that Canada is a constitutional monarchy by looking at the list of elements of CM, I have no idea

[25]This echoes Soames's point, cited at the end of §1, above.

what *makes* Canada a constitutional monarchy, or *what it is* for Canada to be a constitutional monarchy.

Our definition, above, of the set $TRUE_P$ does not proceed simply by giving a list. However, the definition relies on a list to get off the ground. The reason is that, in stipulating that P speaks the language L, you are effectively giving a list-like stipulation of the most basic word-world relation. This relation is given by the interpretation function I_L, which is simply a set of ordered pairs, a list assigning some object to each name, some set of objects to each predicate symbol, and so on. If we want an illuminating specification of the sentences that are true-for-P, then we might want an illuminating account of which nonlinguistic objects correspond to which terms and predicates, as P uses the terms and predicates—here we postpone the issue of in what sense we want the account to be illuminating or what we want it to illuminate. To give such an account presumably requires more than merely stipulating that P speaks language L, which is, in effect the giving of a list: we want to know on account of *what* P speaks language L.

In [3], Field imagines giving a general illuminating definition of "α is true-for-P" in two steps. First we give an illuminating non-list-like account of "the name c refers (for P) to the object o" and "the predicate G refers (for P) to the set S of objects". In particular, Field envisions giving a physicalistically acceptable account of reference-for-P in something like causal terms: in terms of complex causal relations between P's uses of the basic expressions of his language and the objects or sets of objects with which P interacts. Note that it is a contingent question of what object the name c refers-to-for-P. The first part of Field's definition gives us, in effect, an illuminating account of *what language* P *speaks*. The second part of the definition is then to define truth-for-P as truth-in-L where L is the language that P speaks.[26]

[26]In order to give a truly non-list-like definition of truth, we should also give a physicalistically acceptable account of what it is for a speaker to use this expression for disjunction, that expression for universal quantification, and so on. Presumably, the account will *not* be in terms of causal interactions between expressions on one hand and disjunction, universal quantification, or whatever on the other.

The biconditional

(14) ‘Snow is white' is true-for-P iff snow is white,

is now contingent since it is a contingent question whether ‘snow' refers-for-P to snow. If things had been different, P's use of the word ‘snow' might have been causally related to grass, in which case P would have spoken a language other than English. This account of truth for a speaker is a kind of causal-correspondence account, since it essentially involves a causal story about which names correspond to which objects and so on. This story will likely also involve "the thoughts and practices of speakers of language" as McGee and McLaughlin put it.

The last part of Field's story should tell us what this causal-correspondence theory of truth-for-P is good for. He voices his hunch as follows: "the original purpose of the notion of truth ... was to aid us in utilizing the utterances of others in drawing conclusions about the world." A causal-correspondence notion of Truth-for-P ties a speaker's assertions to the actual objects that the assertions are about, in such a way that we can account for that speaker's general reliability regarding these objects. More importantly, we can thus *use* the speaker's assertions to draw conclusion about the world. An account of what language P speaks should be "illuminating" in this sense: it should be part of a story that sheds light on our ability to use others' assertions to draw conclusions about the world.[27]

Can we use the notion of correspondence to help us interpret the fixed point semantics? One thing our discussion has brought out is the potential utility of both a disquotational notion of truth, for such things as semantic ascent, and a notion of correspondence, to account for speakers'

[27]It is worth noting that Field's project is less "bottom-up" than it might seem: the appropriateness of a particular causal account of reference ultimately depends on the utility of the resulting notion of truth for explaining our ability to use the utterances of others to draw conclusions about the world. So, it might be argued, the ultimate units of semantic significance are sentences and not subsentential expressions. In [8], Field considers a number of other uses to which a correspondence notion of truth might be put, or for which it might be needed. In [8], Field is leaning strongly towards a deflationary conception of truth. In [9] and [10], his conversion seems complete.

reliability.[28] Grover [12] develops the disquotational intuition: she suggests that the sentence ‘α is true' is to be understood as an *inheritor*. (See also [13].) An *inheritor* is an expression that somehow picks up its content from another expression. The classic inheritors are pronouns of laziness, as used in ‘Dorothy went to the store and she bought gloves': the pronoun ‘she' picks up it referent from its antecedent. The antecedent of the sentence ‘α is true' is the sentence α, which itself might have an antecedent. If all goes well, these inheritance chains come to an end, and get "grounded" in sentences which are not themselves inheritors, but which get their content in some other way.

Suppose that we combine the idea of the *object-language* predicate T as an inheritance device together with the desire to develop a *metalinguistic* notion not of *truth* but of *correspondence* for some speaker P: a notion of correspondence that will explain the fact that we can *use* P's assertions to learn about the world. Suppose that we have somehow cooked up a non-list-like physicalistically acceptable specification of reference for all of the names and predicates that P uses, other than T. This determines a ground model for P, though which ground model it determines will be an empirical question. Given a ground model for P, can we develop a metalinguistic notion of correspondence-for-P?

Before I try to answer that question, I want to deal with two issues. First, I want to point out one way in which our intuitions concerning correspondence diverge from our intuitions concerning truth-as-an-inheritance-device. With truth so conceived, the claim that a sentence is not true should have the same content as the claim that the negation of the sentence is true. But with *correspondence* this intuition is not so strong. In particular, there seem to be *two* ways in which a sentence can fail to correspond (can fail to be useful for drawing conclusions about the world): on the one hand its negation might correspond; on the other hand it might be semantically defective in some way, for example it might contain a non-referring singular term.

---

[28]In §4, we saw that McGee and McLaughlin [23] also conceive of their metalinguistic notion as a correspondence notion, which they distinguish from a deflationary notion of truth. See also McGee [22]. I will say more about their conception of affairs at the end of this section.

Or, what is more relevant here, it might be an ungrounded inheritor. Given that there are two ways that a sentence can fail to correspond, it seems we want to define, in the metalanguage, three sets of object-language sentences: The set of corresponders; the set of anti-corresponders (these are the sentences whose negations correspond); and the set of ungrounded sentences. Among the corresponders we want the atomic sentences Gc where the referent of c is in the set referred to by G. And we do *not* want inheritors that never get grounded, nor do we want their negations.

Second, I want to point out that the truth-as-inheritance-device intuition suggests the supervenience of semantics. I might have prejudiced the case by supposing that we have cooked up a non-list-like physicalistically acceptable specification of reference for all of the names and predicates that P uses, other than T. But, if we are going to interpret T as an inheritance device, we *already have* an interpretation of T, and we should not have to add another one on top of it. Another way to see the point is to consider a truth teller: suppose that τ is a name, $M = \langle D, I \rangle$ is a ground model representing a community's use of the T-free fragment of its language, and $I(\tau) = T\tau$. On the nonsupervenience approach, the community could be speaking a language L in which $T\tau$ corresponds: i.e. $T\tau$ is the kind of thing we can use as a guide to the world. But there does not seem to be any non-circular place for $T\tau$ to get its content: $T\tau$ initiates an infinitely long inheritance chain, $T\tau$, $T\tau$, $T\tau$, ... It never gets "grounded" in a sentence which is a non-inheritor, and which gets it content in some other way. To the extent that the chain *does* get grounded, it gets grounded in the sentence $T\tau$ itself: $T\tau$ might get its content in some other way—e.g. by stipulation of the community's elders—but in that case T is no longer acting purely as an inheritance device.

My ploy is going to be pretty transparent. I am going to make a go of using the fixed point semantics to define the metalinguistic notions of correspondence, anti-correspondence, and ungroundedness. Starting with a ground model, I am going to settle on a particular fixed point. I am going to suggest a philosophical interpretation of this fixed point: it provides the desired

three-way partition of the sentences. We do *not* have a metalinguistic notion of *truth*: that is, we do not have a metalinguistic predicate ‘x is true' that acts as an inheritance device. But we will have a metalinguistic predicate, ‘x corresponds'. And I suggest that this notion can serve Field's purpose, "to aid us in utilizing the utterances of others in drawing conclusions about the world."

There are two decisions to make, and the current interpretive setting might help us make them. Decision 1: Which scheme should we use for evaluating the correspondence-value of compound sentences? Decision 2: Which fixed point?

We can make Decision 2 by considering Kripke's construction, in [19], of a particular fixed point, the so-called *least* fixed point. Suppose that we have assigned correspondence values to all the sentences that do not contain the predicate T. These sentences will all be grounded. So if α if one of these sentences, then T‘α' will also be grounded and should get the same correspondence value. So this generates correspondence values for a bigger set of sentences. We then use the appropriate scheme to get correspondence values for compounds of these. But we now have an even bigger set of grounded sentences. For each α among these, we can evaluate the sentence T‘α', getting an even bigger set of grounded sentences. This procedure will eventually lead us to a particular fixed point, the *least* fixed point. Kripke defines a sentence to be *grounded* iff it gets a definite truth-value in the least fixed point. In the current setting, we have given the same definition of groundedness, but our motivation has been in terms of the interpretation of the object-language predicate T has an inheritance device, and in terms of the metalinguistically defined properties not being *truth* and *falsity* and *neither*, but *corresponders*, *anticorresponders* and *ungrounded*.[29]

We still have Decision 1: which scheme should we use for evaluating the correspondence value of compound sentences? If we have to choose between the Kleene schemes (adapted to

[29]I do not see how to provide a similar interpretation of the totality of fixed points, or of any other particular fixed point.

correspondence values), the question boils down to this: If P corresponds and Q is ungrounded, does (P $\vee$ Q) correspond or is (P $\vee$ Q) ungrounded? Recall that the idea that an utterance corresponds is ultimately tied to the notion of the *utility* of the utterance for the listeners who are trying to find out about the world. Suppose I know that P corresponds and that Q is ungrounded and that R anti-corresponds, i.e. that $\neg$R corresponds. And suppose that I utter (P $\vee$ Q) on one occasion and (P $\vee$ R) on another. By the adapted classical rules for $\vee$, we say that (P $\vee$ R) corresponds. But when I uttered (P $\vee$ Q), was the *utility* of my utterance to my listeners not just as great as when I uttered (P $\vee$ R)? It seems the utility of both utterances is similar: taken seriously, they both narrow down the options; they both guide my listeners into exploring two possibilities rather than more; and such. So it seems that if we take (P $\vee$ R) to correspond when P corresponds and $\neg$R corresponds, then we should also take (P $\vee$ Q) to correspond when P corresponds and Q is ungrounded. Thus we should opt for the strong Kleene scheme rather than the weak one.

What about the supervaluation scheme, which agrees with the strong Kleene scheme on this example? There are disagreements: according to the strong Kleene scheme, (Q $\vee$ $\neg$Q) is ungrounded, since Q is ungrounded, while according to the supervaluation scheme, (Q $\vee$ $\neg$Q) corresponds. On one intuition, (Q $\vee$ $\neg$Q) is in fact ungrounded: it does not inherit any more content than its disjuncts. Perhaps it gets its content from its syntactic form—all logical theorems are grounded on the supervaluation scheme. But there is a quite different argument against the supervaluation scheme.

In the context of the semantics for vague predicates, we say that a sentence is *definitely true in* L iff it is true in all acceptable precisifications L′ of L. The notion of truth in a precisification L′ is classical since L′ is classical. McGee [22] presents the following gloss on how we should understand an acceptable precisification of L: "A possible world [i.e. an acceptable precisification] represents, not a way things might have been, but a way things might be consistent with the totality of the empirical facts and our linguistic commitments" (pp. 197-8).

(Here the "totality of empirical facts" is represented by the gappy interpretation of the language, and the "linguistic commitments" are represented by the set of constraints.)

This gloss is plausible enough in the context of the semantics for vague predicates: if Harry is a borderline case of baldness, then both the claim that Harry is bald and the claim that Harry is not bald are consistent with the empirical facts and our linguistic commitments, because these facts and commitments leave the question of Harry's baldness open. But in the context of the semantics for languages with their own truth predicate, this gloss does not work. If we try to push the analogy through, then we have to say that the liar is a borderline case of truth. But we do not want to continue by saying that both the claim that the liar is true and the claim that it is not true are consistent with the empirical facts and our linguistic commitments. In fact, each of these claims is *inconsistent* with empirical fact that $I(\lambda) = \neg T\lambda$ and with our linguistic commitments concerning truth.

So the notion of a precisification has a coherent interpretation in the semantics for vagueness that it just does not have in the semantics for languages with their own truth predicates. But the supervaluation scheme is only as philosophically coherent as is the notion of a precisification. So, pending a better gloss on the notion of a precisification, I conclude that we should reject the supervaluation scheme.

I have been exploring the idea that the metalinguistic notion '$x \in TRUE_L$' is not a notion of truth, but a notion of correspondence—a notion useful in explaining why we can draw conclusions about the world from the utterances of others. This is very close to McGee and McLaughlin's bifurcation of the notion of truth into a correspondence notion of definite truth, and a deflationary notion of truth.[30] Shortly, I will contrast the approach in McGee and McLaughlin [23] and McGee [22] to my approach (which was arrived at independently). Before I do, I want to point out that the fact that they think of the two notions as distinct notions of *truth* while I

[30]McGee [22] credits this bifurcation to Field [8].

have declined to use the word ‘truth' for the metalinguistic notion is really only a terminological difference.

The main difference between their approach and mine is that they consider the truth predicate to be on a par, semantically, with the other predicates. This is especially clear in McGee [22]. McGee explicitly proposes treating the object-language predicate T as a vague predicate with a semantic value of the same sort as the semantic value of ‘bald'. To be sure, T has a different semantic value from ‘bald' and the penumbral connections surrounding T are quite different from those surrounding ‘bald', but the two predicates are of the same semantic type. The main problem I see with this assimilation is that, while the supervaluation scheme seems appropriate for the semantics of vagueness, I have argued that it is innappropriate for the semantics of truth.

I have proposed Grover's [12] treatment of T as an inheritance device, or an inheritor-forming operator. On this line, one does not give the semantic contribution that T makes to the sentences T occurs in by displaying an extension and antiextension for T. Rather, the semantic contribution that the occurrence of T in T‘α' makes to T‘α' is to point T‘α' anaphorically to the sentence α, so that they are given the same correspondence conditions. T *appears* to be assigned an extension-antiextension pair in the least fixed point and in each gappy interpreted language leading up to the least fixed point in the Kripkean construction. But the ordered pairs apparently assigned to T as the construction proceeds are better thought of as progressive estimates of the set of corresponders and anti-corresponders, rather than progressive estimates of the semantic value of T.

§7. Conclusion. A semantics for a language is often given by assigning objects to expressions, and giving rules according to which the semantic values of compound expressions depends in some systematic way on the semantics values of their parts. When presented with a semantics, it is always legitimate to ask, what is the significance of the semantic notions thus introduced?

The semantic value of declarative sentences will often be truth values, or some other objects which determine truth values. This need not be the case: one might, for example, hope for a semantics that assigns to each sentence the mode of its verification, or some idealization of it.

I have been taking Tarski's semantics to partition the sentences into two subsets, TRUE and FALSE. This is equivalent to assigning to each sentence one of two truth values, T or F. I have considered one answer as to the significance of Tarski's semantics: because of the T-biconditionals, we can use the technical notion of truth to effect the semantic ascent wanted by the logical positivists. Tarski's semantics might be even more significant than that: with the appropriate supplementation, we might be able to use Tarski's semantic notions to shed light on language acquisition, the content of thoughts, the utility of others' utterances in drawing conclusions about the world, and so on.

The fixed point semantics also seems to assign truth values to sentences—perhaps not to *all* sentences, but to many of them. (If a sentence $\alpha \in \text{NEITHER}_L$, then, on a natural account, α is assigned no truth value at all.) But I have argued that the notion of "truth" introduced into the metalanguage by the fixed point semantic theorist—as opposed to the notion of truth in the object language—is not genuinely a notion of truth. And it becomes legitimate to ask, what is the point of the notion? What purpose can it serve?

One possibility is to adapt Field's suggestion that, with the appropriate supplementation, the classical notion of truth can be used in the explanation of why we can draw conclusions about the world from utterances of others. Field suggests a causal-correspondence theory of truth, based on a causal-correspondence theory of reference together with a Tarskian definition of truth in terms of reference. I explored the possibility of putting the fixed point semantics' ersatz notion of "truth"—what I simply call *correspondence*—to the same task as Field wants to put the classical notion of truth. One bonus of this approach is that it gives us principled reasons for choosing among fixed points, and for choosing among schemes for assigning values to composite sentences.

Unfortunately, causal-correspondence theories—whether of truth or of *correspondence*—are endangered by a number of difficulties. There is the inscrutability of reference: it is simply unclear that we can tell a causal story picking out a determinate reference for each name and each predicate, so as to explain the reliability of speakers.[31] There is the problem of reference to abstract objects, such as numbers, which can hardly be accounted for in causal terms.[32] There is a further worry, expressed by Field in [9] that bears directly on his proposal in [3]: "there are plenty of examples where the indication relations [the supposed causal-physical correlations between between belief states and utterances on one hand and the world on the other] don't reflect what we would intuitively regard as truth conditions."

Unhappy with correspondence theories, we might look for other interpretations of the metalinguistic notions produced by the fixed point semantics. Often contrasted to correspondence theories of truth is Davidson's distinctive appropriation of Tarski's work. For Davidson, the role of the metalinguistic semantic notion of truth is the part it plays in *interpreting* others. To interpret another is to try to make sense of her as a rational agent. This involves two interwoven projects: to attribute to her beliefs and desires, and to provide for her utterances a Tarskian theory of truth. An interpretation is successful insofar as it succeeds in rendering the other as rational, and this involves rendering true most of her beliefs, desires and utterances. There is no need to provide a causal-correspondence theory or any other theory of the reference of an interpretee's subsentential expressions, beyond a merely stipulative list. As long as the holistic constraints are met, the interpretation will be reckoned as successful.

I think that it would be an extremely interesting project to consider whether a Davidsonian interpreter can interpret a speaker who has a truth predicate that she applies to her own sentences. If the interpreter is constrained to using a Tarskian theory of truth, then no, not if the interpretee

---

[31]In [4], [5] and [6], Field argues quite convincingly that the inscrutability of reference can be handled within an appropriately tweaked version of a correspondence theory of truth.

[32]This, I take it, constitutes some of the motivation for [7].

can express the liar's paradox. But perhaps the interpreter can appeal to a fixed point semantics, or some other Tarski-inspired semantics for languages with their own truth predicates.

Fixed point semantics are not the only semantics for languages with their own truth predicates. Gupta and Belnap's [14] revision theory of truth, for example, is like the fixed point semantics in that it builds upon the classical neo-Tarskian notion of truth-in-a-model.[33] Gupta and Belnap never define "α is true in M", for sentences α and ground models M. But they do define various notions of "α is valid in M", all of which agree with the neo-Tarskian definition of "α is true in M" in non-paradoxical contexts. It would be worth considering what the significance is of these notions of validity-in-M, along the same lines I have been pursuing in this paper.

Acknowledgements. Thanks to an audience at UC Irvine, particularly Peter Woodruff, and an audience at McMaster University for helpful comments. Thanks also to Greg Restall for directing me to Hodges's paper on truth in a structure. Thanks especially to Graham Priest. In a brief conversation, I suggested that the metalinguistic notion of truth in fixed point semantics might not actually be a notion of *truth*. He asked, "Then in what sense is it a *semantics*?"

REFERENCES

[1] L. Davis 1979, "An alternate formulation of Kripke's theory of truth", *Journal of Philosophical Logic* 8, 289-296.
[2] J. Etchemendy 1988, "Tarski on truth and logical consequence", *Journal of Symbolic Logic* 53, 51-79.
[3] H. Field 1972, "Tarski's theory of truth", *Journal of Philosophy* 69, 347-75.
[4] H. Field 1973, "Theory change and the indeterminacy of reference", *Journal of Philosophy* 70.
[5] H. Field 1974, "Quine and correspondence theory", *Philosophical Review* 83, 200-28.
[6] H. Field 1975, "Conventionalism and instrumentalism in semantics", *Noûs* 9.
[7] H. Field 1980, *Science Without Numbers: A Defence of Nominalism*, Princeton University Press, Princeton NJ.
[8] H. Field 1986, "The deflationary conception of truth", in *Fact, Science and Morality: Essays on A.J. Ayer's* Language, Truth and Logic (G. Macdonald and C. Wright, eds.), Blackwell, London, 55-117.
[9] H. Field 1994, "Deflationist views of meaning and content", *Mind* 103, 249-85.
[10] H. Field 1994, "Disquotational truth and factually defective discourse", *Philosophical Review* 103, 405-52.
[11] K. Fine 1975, "Vagueness, Truth, and Logic", *Synthese* 30, 265-300.
[12] D. Grover 1977, "Inheritors and paradox", *Journal of Philosophy* 74, 590-604.

[33]Gupta and Belnap point out that the notion of truth they are modelling is not truth-in-a-model or truth-in-an-interpreted-language but "absolute" truth. They define metalinguistic semantic notions in order to give a theory of an object-language notion of absolute truth. It is the metalinguistic notions which are variants on the classical neo-Tarskian notion of truth-in-a-model.

[13] D. Grover, J. Camp and N. Belnap 1975, "A prosentential theory of truth", *Philosophical Studies* 27, 73-125.
[14] A. Gupta and N. Belnap 1993, *The Revision Theory of Truth*, MIT Press.
[15] S. Haack 1978, *Philosophy of Logics*, Cambridge University Press.
[16] W. Hodges 1986, "Truth in a structure", *Proceedings of the Aristotelian Society* n.s. 86, 135-51.
[17] R. Kirkham 1992, *Theories of Truth: A Critical Introduction*, MIT Press.
[18] M. Kremer 1988, "Kripke and the logic of truth", *Journal of Philosophical Logic*, 225-78.
[19] S. Kripke 1975, "Outline of a theory of truth", *Journal of Philosophy*, 690-716.
[20] F. Kroon 1984, "Steinus on the paradoxes", *Theoria* 50, 178-211.
[21] R.L. Martin and P.W. Woodruff 1975, "On representing 'True-in-L' in L", *Philosophia* 5, 217-21.
[22] V. McGee 1991, *Truth, Vagueness and Paradox*, Hackett, Indianapolis.
[23] V. McGee and B. McLaughlin 1994, "Distinctions without a difference", *Southern Journal of Philosophy* 33 (supp), 203-51.
[24] T. Parsons 1984, "Assertion, denial and the liar paradox", *Journal of Philosophical Logic* 13, 137-152.
[25] S. Read 1994, *Thinking About Logic: An Introduction to the Philosophy of Logic*, Oxford University Press.
[26] S. Soames 1984, "What is a theory of truth?" *Journal of Philosophy* 81, 411-29.
[27] A. Tarski 1933, *Pojęcie Prawdy w Językach nauk Dedukcyjnych* (On the concept of truth in languages of deductive sciences), Warsaw, 1933. English translation, "The concept of truth in formalized languages", in Tarski 1983, 115-278.
[28] A. Tarski 1935, "Der Wahrheitsbegriff in den formalisierten Sprachen", *Studia Philosophica I*, 261-405. German translation of Tarski 1933.
[29] A. Tarski 1936, "O ugruntowaniu naukowej semantyki", *Przeglad Filozoficzny* 39, 50-7. English translation, "The establishment of scientific semantics", in Tarski 1983, 401-8.
[30] A. Tarski 1944, "The semantic conception of truth and the foundations of semantics", *Philosophy and Phenomenological Research* 4, 341-76.
[31] A. Tarski 1952, "Some notions and methods on the borderline of algebra and metamathematics", in *Proceedings of the International Congress of Mathematicians, Cambridge, Mass., 1950, Volume 1*, American Mathematical Society, Providence RI, 705-20.
[32] A. Tarski 1983, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938* (J.H. Woodger, trans.), 2nd edition, Hackett, Indianapolis.
[33] A. Tarski and R.L. Vaught 1957, "Arithmetical extensions of relational systems", *Compositio Mathematica* 13, 81-102.
[34] B. van Fraassen 1968, "Presupposition, implication and self-reference", *Journal of Philosophy* 65, 136-52.