Axiomatizing Truth: Why and How?

Solomon Feferman*

For Helmut Schwichtenberg on the occasion of his 70th birthday

Broadly speaking there are two kinds of theories of truth, philosophical and logical. The philosophical theories of truth go back to the Greeks and forward to the present day. Among these are the correspondence, coherentist, pragmatist, deflationary and primitivist theories of truth. Logical theories of truth are roughly of two kinds, semantical (or definitional) and axiomatic. Tarski inaugurated semantical theories in the mid 1930s with his definition of truth for a logically circumscribed language within a metalanguage for it, i.e. in a typed setting in order to avoid inconsistency. However, the ordinary use of truth in natural language is untyped, and so beginning in the 1960s, attempts were made to obtain useful consistent untyped semantical theories by giving up some of Tarski's basic assumptions. One of the most successful was due to Kripke in 1975, who defined a notion of truth for an untyped three-valued language. My axiomatization of Kripke's model a few years later inaugurated a considerable body of work on a variety of axiomatic theories that continues to be actively pursued. In this paper general considerations are presented as to why one should axiomatize theories of truth and what criteria should be applied to them. These are then illustrated with three examples from my own work as to how one might try to go about meeting these criteria.

1 Introduction

Broadly speaking there are two kinds of theories of truth, philosophical and logical. The philosophical theories of truth go back to the Greeks and forward to the present day, including the correspondence, coherentist, pragmatist, deflationary and primitivist theories of truth.[1] Logical theories of truth, on the other hand, date only

*This paper is based on an invited lecture that I gave for the Pillars of Truth conference held at Princeton University, April 8–10, 2011.

[1] As sources for these, both Kirkham (1995) and Burgess and Burgess (2011) provide excellent surveys and expositions, while there are several fine collections of original articles such as those of Blackburn and Simmons (1999) and Lynch (2001).

to the 1930s. They are roughly of two kinds, semantical (or definitional) and axiomatic. Tarski (1935) inaugurated semantical theories with his definition of truth for a logically circumscribed language within a metalanguage for it, i.e. in a typed setting. He argued that this was necessary since a language which contains its own truth predicate is inconsistent if it satisfies a few basic assumptions, namely the T-scheme, classical propositional logic,[2] and the capacity to form self-referential statements. However, the ordinary use of truth in natural language is untyped and the constraints of a hierarchical theory seem unduly restrictive. Moreover, mathematics provided an excellent example in replacing the theory of types by untyped systems of set theory. Thus it was that beginning in the 1960s, attempts were made to obtain useful consistent untyped semantical theories of truth by giving up some part of Tarski's basic assumptions.[3] One of the most influential of these was that due to Kripke (1975), who defined a notion of truth for an untyped three-valued language.[4] My axiomatization of Kripke's model a few years later (see Feferman 1991) inaugurated a considerable body of work on a variety of axiomatic theories that continues to be actively pursued.[5] In this paper general considerations are presented as to why one should axiomatize theories of truth and what criteria ought to be applied to them. These are then illustrated with three examples from my own work (because that is what I know best) as to how one might try to go about meeting these criteria and to what extent one may succeed.

2 Why axiomatize theories of truth?[6]

1. Axiomatic theories separate out what is needed to justify a given semantical definition from what the definition is designed to achieve. The axiomatic theory is usually much weaker than the (implicit) ambient theory in which the semantical construction is carried out. Often the latter includes a considerable portion of set theory.

[2] Actually, intuitionistic logic suffices to derive a contradiction.

[3] A number of these efforts can be found in the articles collected in Martin (1970) and Martin (1984).

[4] Martin and Woodruff (1975) independently arrived at a closely related model. Theirs produces a maximal fixed point for a certain monotonic operator, whereas Kripke obtained minimal fixed points among others. Among other semantic approaches are those due to Barwise and Etchemendy (1987) and Gupta and Belnap (1993).

[5] The first book to exposit a number of axiomatic theories was Cantini (1996); much of that has been brought up to date in the excellent expository work, Halbach (2011); both incorporate original research by the respective authors. Some of the noteworthy books devoted to specific axiomatic approaches are those due to McGee (1991), Maudlin (2004), Field (2008) and Horsten (2011).

[6] There is some overlap in this section with the motivations for axiomatization given in Ch. 1 of Halbach (2011).

2. Many axiomatic theories of truth are based on a given semantic definition and are thus automatically consistent. But those that are not may be shown consistent by providing a suitable model or by proof-theoretical means.

3. Often, particular semantic constructions are designed to realize certain *prima facie* required basic properties of truth such as the T-scheme or compositionality. What axiomatizing such a construction shows is precisely the extent to which such properties are met.

4. Moreover, such axiomatization provides the explicit statement of *further* properties that were not necessarily part of the original motivations for the construction.

5. An axiomatization (not necessarily uniquely determined) of a given semantical construction provides a framework in which one can reason systematically about various aspects of the construction. This helps to assess the value and possible defects of such a construction.

6. One can compare like and unlike axiomatizations as to their proof-theoretical strength using an extensive body of well-established metamathematical techniques.

7. Given axiomatizations suggest natural variants such as by extending general principles from one's base theory (e.g., induction in arithmetic, or separation in set theory) to the theory with a truth predicate.

8. A given philosophical conception of truth may suggest a semantical construction or an axiomatization, and once made more explicit in the latter way, we are in a better position to assess the underlying conception.

It should be remarked that not all philosophical theories of truth are amenable to axiomatic representation. How axiomatize, for example, coherence or pragmatic theories of truth?

3 Criteria for a theory of truth

Various criteria have been proposed for consistent axiomatic theories of truth that contain their own truth predicate. To my mind, the best articulation of such criteria is one due to Hannes Leitgeb (2007).[7] He sets down eight of these, each of

[7] But see also Sheard (2002). Another interesting discussion of criteria is to be found in McDonald (2000).

which has plausibility in its own right, and various of which are ordinarily taken for granted, but the combination of all of which cannot simultaneously be realized on pain of inconsistency. That is why Leitgeb calls his piece, "What theories of truth should be like (but cannot be)". His eight criteria (L1-L8) are as follows.

(L1) Truth should be expressed by a predicate (and a theory of syntax should be available).

(L2) If a theory of truth is added to mathematical or empirical theories, it should be possible to prove the latter true.

(L3) The truth predicate should not be subject to any type restrictions.

(L4) T-biconditionals [in the T-scheme] should be derivable unrestrictedly.

(L5) Truth should be compositional.

(L6) The theory should allow for standard interpretations.

(L7) The outer logic and the inner logic should coincide.

(L8) The outer logic should be classical.

These can be spelled out more precisely, as follows.

To meet (L1), we have first to address the common philosophical issue whether truth is a predicate of sentences or of propositions. We would certainly grant that if two sentences, from the same or different languages, express the same proposition then their truth conditions agree. That would seem to argue in favor of truth as a predicate of propositions. The argument in favor of sentences is that we have excellent theories of sentences as structured syntactic objects; these can be dealt with in full precision and with great flexibility in formal theories of syntax as provided for example in concatenation theory, or elementary set theory, or in arithmetic via Gödel coding. The nature of propositions is obscure by comparison; one issue is whether or not they are structured objects. And what does it mean for a sentence to express a proposition? When do two sentences express the same proposition? Are all propositions expressible in some language? Finally, do all sentences in a given language express a proposition? When we settle, as is customary in work on axiomatic theories of truth, on sentences being the truth-bearers, one avoids dealing with all but the last of these difficult questions, and concentrates instead in each axiomatic theory on a more precise question: Which sentences are taken to be the truth bearers? [8]

[8] Some other arguments in favor of sentences as the truth-bearers can be found in Halbach (2011), pp. 9–14, and Horsten (2011), pp. 2–3.

Given that we will take the truth predicate to be applicable to some or all sentences within a specified formal language, it is standard in the preponderance of the literature to take Peano Arithmetic (PA) as a base theory, though weaker theories suffice for much of the work. Let L be the language of PA and L_T be its extension by the unary predicate symbol $T(x)$. We shall be considering formal systems S extending PA whose language $L(S)$ includes L_T. We use $A, B, C, \ldots$ to range over formulas and sentences of $L(S)$. Given a sentence A of $L(S)$, let $\#A$ be the numeral of the Gödel number of A, so we can write $T(\#A)$ to express that A is true; for simplicity in the following, write $T(A)$ for $T(\#A)$. Among other things, this choice of coding means that we can apply Gödel's method of constructing self-referential sentences. That is, given any formula $C(x)$ of $L(S)$ we can construct a sentence A such that $A \leftrightarrow C(A)$ is provable in S. In particular, we can construct the "Liar" sentence using $\neg T(x)$ for $C(x)$; I'll use Λ to denote it instead of a Roman cap letter, i.e. Λ is a sentence of L_T and S proves $\Lambda \leftrightarrow \neg T(\Lambda)$.[9]

Now, a minimum requirement for (L2) is that S proves the sentence P: "all sentences provable in PA are true".

For (L3), $T(A)$ is syntactically acceptable for every sentence A of $L(S)$.[10]

By the T-biconditionals in (L4) is meant the sentences $T(A) \leftrightarrow A$ for A in $L(S)$. Thus (L4) requires that all of these are provable in S.

Compositionality (L5) means that S proves $T(\neg A) \leftrightarrow \neg T(A)$, $T(A \vee B) \leftrightarrow T(A) \vee T(B)$, $T(\forall x A(x)) \leftrightarrow \forall x T(A(\mathrm{num}.(x)))$, and so on for the other propositional operators and the existential quantifier. To formulate more general versions, we use operations on numbers that correspond to the logical operations, via the "dot" notation, i.e. $\neg.$, $\vee.$,$\forall.$, etc. Thus, for example $\neg.\#A = \#(\neg A)$. Then we can write $\forall x(\mathrm{Sent}_{L(S)}(x) \rightarrow [T(\neg.x) \leftrightarrow \neg T(x)])$, and so on, where $\mathrm{Sent}_{L(S)}(x)$ is a formula of L expressing that x is the number of a sentence of $L(S)$.

(L6) means that S has a model in which the language of PA is given its standard interpretation.

For (L7), by the "outer logic" of S is meant its basic logical axioms and rules. By its "inner logic" is meant the laws holding for those A such that S proves $T(A)$.

(L8) simply means that the basic logic of S (its "outer logic") is classical.

Tarski's Undefinability Theorem. If S satisfies (L1), (L3), (L4) and (L8) then S is inconsistent.

The proof as usual makes use from (L1) and (L3) of the construction of a Liar sentence Λ such that S proves $\Lambda \leftrightarrow \neg T(\Lambda)$, which when combined with the as-

[9] Often, 'λ' is used for a Liar sentence.

[10] Actually, as long as the truth predicate T is a predicate of numbers, we would say that $T(A)$ is syntactically acceptable for every sentence A (via its Gödel number) even for a typed theory of truth. The significance of (L3) lies in connection with (L4) which would be restricted in the case of a typed theory; the same holds for the compositionality conditions on truth in (L5).

sumption from (L4) that $\Lambda \leftrightarrow T(\Lambda)$, leads to a contradiction under the assumption in (L8) of classical propositional logic. (In fact, as noted in fn. 2 intuitionistic logic suffices.)

The question in view of Tarski's theorem is, if one is to axiomatize truth in a consistent type-free way, which of the criteria to accept and which to reject. Of course one will also want to give arguments for these. Given the aim, one will certainly accept (L1), that we are working within an extension S of PA, and (L3), that all sentences of $L(S)$ are admissible arguments for the T predicate. Moreover, if S is to be consistent, we shall certainly want it to have a model, and it is desirable then to grant (L6) that it has a model which is standard for the natural numbers and hence in which all axioms of PA are true. It is then reasonable to demand that one be able to prove that in S, i.e. to accept (L2).

This leaves (L4), (L5), (L7) and (L8) in question. Let me begin with the last of these, namely that the outer logic of S should be classical. There have been a number of approaches to the consistent type-free axiomatization of theories of truth that are based on a restriction of classical logic, such as one form or another of three-valued logic or many-valued logics more generally, or the rejection of such principles as *ex falso quodlibet* in paraconsistent logics. In Feferman (1984), of the former I wrote that "nothing like sustained ordinary reasoning" can be carried out in them, and it is my impression that the same holds as well for the latter. Other approaches have been based on extensions by new connectives, such as a new conditional or biconditional, which do not satisfy the same laws as the classical (or even intuitionistic) conditional or biconditional, in terms of which the T-scheme is rewritten. One such is given in the third example of axiomatization from my own work below, as an extension of classical logic. I defend such uses, but not those which make essential restrictions in the logic otherwise, such as in Field (2008). So in that sense, I strongly favor (L8). But for the cases where the T-scheme (L4) is not written in terms of a new biconditional, this means that I can only accept it with some restrictions; that in turn may affect how much of compositionality (L5) is accepted. Finally, I reject the demand that the outer logic equal the inner logic (L7), if the inner logic is some essential weakening of classical logic such as those mentioned above.

4 How to axiomatize: Three examples.

These three examples illustrate how my own choices as to which criteria to accept and which to modify or reject have been dealt with in my own work on the axiomatization of truth. The reader is referred in each case to the publication in question for the actual presentation of the systems involved; only certain specific aspects of

those are discussed here as are needed to explain how they relate to the preceding criteria. For certain reasons, I present these in reverse order from that of date of publication.

4.1 Axioms for determinateness and truth (Feferman 2008).

This begins (pp. 206–207) with the statement of a general philosophical position: Every predicate has a domain of significance; it makes sense to apply the predicate only to objects in that domain (cf. also Russell 1908). In the case of the truth predicate T, the domain D in question is taken to consist of the sentences that are meaningful and determinate, i.e. have a definite truth value, true or false.[11] D includes various but not necessarily all grammatically correct sentences that involve the notion of truth itself, for example the statement P that each sentence provable in PA is true. At any rate, $T(A) \to D(A)$ holds for each sentence A. Write $F(A)$ for $T(\neg A)$; then also $F(A) \to D(A)$ for each A. It follows that $D(A)$ can here be defined as $T(A) \vee F(A)$, but there is a reason for keeping it as an additional basic predicate, namely to state conditions on it that are prior to those for T. Thus, in the case of this example, $L(S)$ is the extension of L by T and D, and S itself is denoted DT. In accordance with the preceding, (L1), (L3) and (L8) are accepted but (L4) and (L5) need to be restricted. Namely, the restriction of the T-scheme is to take the form,

(L4)* $D(A) \to (T(A) \leftrightarrow A)$ for each sentence A in $L(DT)$.

Similarly, compositionality for T should hold only under assumption of D for all the formulas involved. The basic logical operations of DT are $\neg$, $\vee$, $\to$, and $\forall$, with $\to$ not defined in terms of $\neg$ and $\vee$, but its logic is the standard one of the classical conditional, so that (L8) holds in full. Every L sentence satisfies D. The (strong) compositionality axioms for both D and T come in pairs and are expressed in generality via variables that are taken to range over the formulas of $L(DT)$. For example, we assume

$$D(x \vee .y) \leftrightarrow D(x) \wedge D(y) ,$$

i.e. D holds of a disjunction iff it holds of both disjuncts, and then as usual,

$$D(x \vee .y) \to [T(x \vee .y) \leftrightarrow T(x) \vee T(y)] .$$

The axioms for negation and universal quantification are treated similarly. However, $A \to B$ behaves differently from $\neg A \vee B$ within the context of the D predicate

[11] Some authors, e.g. Kripke (1975) regard the Liar sentence as meaningful though not determinate. I do not agree, but in order to avoid controversy on this point, allow for the possibility of meaningful statements that are not determinate.

for a technical reason to be explained below. Namely, we take as axiom

$$D(x \to .y) \leftrightarrow D(x) \wedge [T(x) \to D(y)] \,,$$

but we still assume as usual

$$D(x \to .y) \to [T(x \to .y) \leftrightarrow (T(x) \to T(y))] \,.$$

The restricted T-scheme (L4)* above is then proved as a direct consequence of these axioms and we have full compositionality for T under D, i.e. a form (L5)* of (L5). Coming back to the D axiom for the conditional, it turns out that if $\to$ were defined as usual in terms of $\neg$ and $\vee$, we could prove the sentence

$$P := \forall x[\mathrm{Sent}_L(x) \wedge \mathrm{Prov}_{PA}(x) \to T(x)].$$

This may be enough to satisfy (L2), depending on how we interpret it. A stronger reading of (L2) is that we should also be able to prove $T(P)$ as well, and for that we need the above condition on D for conditionals.

The consistency of DT is proved by exhibiting a model which is standard for the natural numbers, so that (L6) is met. Moreover, we have $D(A)$ for each sentence A of the language L of arithmetic, so we have the standard truth conditions for those sentences.

Let us turn finally to (L7) and the relation of the outer logic to the inner logic in this system. Constructing a Liar sentence Λ as usual, i.e. one for which $\Lambda \leftrightarrow \neg T(\Lambda)$ is provable, we see that $\neg D(\Lambda)$ must hold, otherwise we would derive a contradiction from (L4)*. It follows that $\neg D(\Lambda \vee \neg\Lambda)$ holds, by our axioms for D on disjunctions. Since $T(\Lambda \vee \neg\Lambda)$ implies $D(\Lambda \vee \neg\Lambda)$, we then also have $\neg T(\Lambda \vee \neg\Lambda)$. But in DT we of course prove $\Lambda \vee \neg\Lambda$, so the outer logic does not equal the inner logic in this case.[12] The situation here is similar in this respect to what was met in the system KF, discussed next.

4.2 The system KF (Feferman 1991).

My axiomatization of the (minimal) three-valued fixed point construction in Kripke (1975) was circulated as notes in 1979. Reinhardt (1985, 1986) took that up for consideration and dubbed the system "Kripke-Feferman", which has since stuck with the abbreviation KF. It then played a central role in the work of McGee (1991) on the axiomatization of truth. The purpose of my own publication involving KF, "Reflecting on incompleteness" (Feferman 1991) was instead completely instrumental, namely to use an axiomatic type-free theory of truth in order to establish

[12] The situation is similar for the sentence called the "Revenge of the Liar."

transfinitely iterated reflection principles without requiring a transfinite hierarchy of truth theories. KF was the basic system introduced for that purpose; it was used there to define the notion of *reflective closure of a schematic axiom system*. However, that was subsequently replaced by a more general notion of *unfolding of a schematic system*, which did not make use of a theory of truth; cf. Feferman (1996) and Feferman and Strahm (2000, 2010).

Since KF took on a life of its own within the work on axiomatic theories of truth, let me consider only one problem about it that has been given much attention, namely that it violates criterion (L7) according to which the outer logic and the inner logic of truth should coincide. In the KF case, the outer logic is classical and the inner one is Kleene's strong three valued logic. Despite the *prima facie* plausible arguments made by Leitgeb, Halbach, Horsten, and others for criterion (L7) I have several reasons why I reject it in this and similar cases; it is not necessary to know the details of the system KF to understand these.

(i) First of all, the distinction between outer and inner logics is only a problem if one conflates two notions of truth, namely the notion of *grounded truth* given by Kripke's least fixed-point construction, and our everyday notion of truth not tied to any particular semantical construction or theory. Thus, in KF, $T(A)$ expresses that the sentence A is a grounded truth while A itself, if provable, is counted as true in the informal sense. So on that reading there is no conflict between accepting both $\neg T(\Lambda \vee \neg\Lambda)$ and $\Lambda \vee \neg\Lambda$ for a formal liar sentence Λ.

(ii) For me, the main direct use of KF is to reason systematically about the properties of the Kripke construction under the *Why* purposes 4 and 5 above. But as I have written in Feferman (1984) p. 264 concerning Kleene's and Lukasiewicz' three valued logic, "nothing like sustained ordinary reasoning can be carried out in either logic." I admire the success of Halbach and Horsten (2006) in axiomatizing the Kripke construction in Kleene 3-valued logic in a system PKF, but inspection of the result has given me all the more reason to disagree with the criterion (L7). In addition, even though Halbach (2011) sec. 16.1 presents a variant of PKF, all the other systems he deals with in his book are based on classical logic, and in Ch. 20 he gives at length a number of reasons for preferring classical systems even where those violate (L7).

(iii) Still, examples like that above with the Liar sentence Λ, or related sentences like the Revenge of the Liar, may give one pause. As to this, I wrote toward the end of (Feferman 2008) where we met the similar problem:

> [T]he provability in DT of sentences $\neg T(A)$ for which A is provable might be regarded as "unintended consequences" or "anomalies" or "little monsters". In a way, this is analogous to other situations in mathematics. For example, to develop a good theory of integration, Lebesgue introduced his theory of measure; that has many excellent properties but also the unintended consequence that there are nonmeasurable sets; [however, their] existence does not affect the positive applications of the theory .... Another example is the existence of space-filling curves as a consequence of a good theory of continuous mappings formulated in purely topological terms.

(iv) One of the reasons (number 6) given for axiomatization at the beginning of this article is that one can compare like and unlike axiomatizations as to their proof-theoretic strength. In Feferman (1991) I introduced two extensions of KF for the notion of reflective closure, $\mathrm{Ref}(PA)$ and $\mathrm{Ref}^*(PA)$, and determined their strengths to be the same as that of the union of the ramified systems RA_α for $\alpha < \varepsilon_0$ and $\alpha < \Gamma_0$, resp. At the end of Feferman (2008) I conjectured that one would obtain the same strengths for the systems DT and a suitable extension DT^*, resp. These conjectures were subsequently verified in Fujimoto (2010). By contrast, the system PKF is relatively weak, as shown in Halbach (2011) sec. 16.2, namely its proof-theoretic strength is the same as the union of the RA_α for $\alpha < \omega^\omega$.[13]

4.3 An axiomatization of deflationism using an intensional equivalence operator (from Feferman 1984).[14]

Deflationism is one of the most popular theories of truth these days. Actually, as explained in Ch. 3 of Burgess and Burgess (2011), this has been spelled out in a great variety of ways, starting with the so-called *redundancy theory* of Ramsey (1927), according to which there is nothing more to the assertion of truth of a sentence than the assertion of the sentence itself. One of the foremost recent proponents of deflationism is Horwich (1990) under the label *minimalism*. Aside from the fact that

[13] In Feferman and Strahm (2000) it was shown that the strength of the full unfolding $U^*(NFA)$ of a basic schematic system NFA of non-finitist arithmetic is the same as that of the union of the RA_α for $\alpha < \Gamma_0$, the least impredicative ordinal.

[14] The article Feferman (1984) was presented as Part I of a two-part article, with the second part to be devoted of the formalisms developed to applications to mathematics, especially the foundations of category theory in an unrestricted sense. But the second part was never written, because the applications did not work out as hoped; instead I have pursued quite different approaches to the foundations of category theory.

Horwich treats truth as a predicate of propositions, rather than sentences, his view is that

> for one to understand the truth predicate is for one to have the disposition to accept any T-biconditional proposition... Against a background of classical logic, this is more or less the same as having the disposition to infer the conclusion proposition from the premise proposition in any T-introduction or T-elimination... (Burgess and Burgess 2011, p. 44)

Considered axiomatically and taking the language to include the truth predicate as well as using sentences rather than propositions, these are actually two different ideas. The first is that one accepts all T-biconditionals in the language of L_T, i.e. all equivalences $T(A) \leftrightarrow A$. The second is that one accepts all inference rules of the form $A/T(A)$ and $T(A)/A$. These are quite different since as we know, over Peano arithmetic in classical logic, the set of T-biconditionals is inconsistent. On the other hand, Friedman and Sheard (1987) have shown that one can consistently accept both all T-Introduction rules and all T-Elimination rules; cf. also Halbach (2011), Ch. 14.

I shall here interpret deflationism in the form that, *by definition*, each $T(A)$ evaluates out to truth (**t**) or falsity (**f**) in the same way as A, allowing for the possibility that both A and $T(A)$ may lack a truth value. Equivalence or equality by definition is taken to be a new connective $\equiv$ different from the truth-functional biconditional $\leftrightarrow$, applicable to instances where both sides may fail to be defined.[15] So, in general, $A \equiv B$ is informally taken to mean that A and B evaluate out to truth or falsity in the same way *when defined*.

In the following, a system of axioms S formulated within the classical first order predicate calculus with equality is presented in the language L of PA extended by the unary predicate symbol T and with the binary sentential operation symbol applicable to all pairs of formulas A, B of $L(S)$ to form $A \equiv B$. In this case we write **t** for the formula $(0 = 0)$ and **f** for its negation, and then take $D(A) =_{\text{def}} (A \equiv \mathbf{t} \vee A \equiv \mathbf{f})$; A is called *definite* (or *determinate*) if $D(A)$ holds. For simplicity, several of the axioms of S are stated informally to encompass a number of formal statements. In the following, unless otherwise specified 'A' is taken to range over the formulas of $L(S)$. For $A(x, y, \dots)$ with possible free variables $x, y, \dots$, $T(A)$ is written for $T(A(\text{num}.x, \text{num}.y, \dots))$. In addition to the axioms of PA, S has the

[15] This is a frequent situation in analysis, for example where one takes

$$f'(x) =_{\text{def}} \lim_{u \to 0}(f(x+u) - f(x)/u)\,.$$

following axioms for $\equiv$, T and D.[16]

Ax. 1 $T(A) \equiv A$

Ax. 2 $\equiv$ is an equivalence relation

Ax. 3 $\neg(\mathbf{t} \equiv \mathbf{f})$

Ax. 4 $\equiv$ preserves $\neg$, $\vee$, $\equiv$ and $\forall$

Ax. 5 $D(A)$ holds for each atomic formula A of L

Ax. 6 D is closed under $\neg$, $\vee$,and $\forall$

Ax. 7 $[A \equiv \mathbf{t} \to A] \wedge [A \equiv \mathbf{f} \to \neg A]$, for each A.

Theorem (Aczel and Feferman 1980, Feferman 1984)**.** *S is a conservative extension of PA.*

NB. In Aczel and Feferman (1980), we wrote $T(A)$ for $A \equiv \mathbf{t}$, while here $T(x)$ is a basic predicate. Ax. 1 replaces the Abstraction Principle (AP) $y \in \{x : A(x)\} \equiv A(y)$ used there; that is essentially the same as the scheme $(T_0)_{\equiv}$ of Feferman (1984), p.268. Note also that a more general theorem is proved op. cit. using S as the extension by the axioms 1-6 of any given extension S_0 of PA in the language of PA.

There are two proofs of this theorem. The first, due to me in the 1980 article with Aczel, makes use of a combinatory style reduction relation for formulas, $A \geq B$, which is shown to satisfy the Church-Rosser theorem. An N-standard model for S is defined in which one takes $A \equiv B$ to hold just in case there exists a C such that $A \geq C$ and $B \geq C$. The second proof, due to Aczel, and presented in Feferman (1984), pp. 268-269, is carried out by turning the 3-valued model of Kripke (1975) into a 2-valued model in an unexpected way.

NB. The connective $\equiv$ fails to satisfy some expected laws such as to infer B from $A \equiv B$ and A. For example, if we take a Liar sentence Λ such that S proves $\Lambda \leftrightarrow \neg T(\Lambda)$, we have $T(\Lambda) \equiv \Lambda$ by Ax. 1, and thus $\neg T(\Lambda) \equiv T(\Lambda)$ by Ax. 2. But $\neg T(\Lambda)$ is true in the just indicated model of S, so if the rule held with $A = \neg T(\Lambda)$ and $B = T(\Lambda)$, we would have a contradiction.

We next show that in combination with a few of the other axioms Ax. 1 leads to the usual truth biconditionals for definite formulas.

[16] I am indebted to Kentaro Fujimoto for his suggestions to improve the formulation of S and its consequences that had been given in a draft of this paper.

Lemma 1. *$D(T(A)) \leftrightarrow D(A)$ for each formula A of $L(S)$.*

Proof. By Ax. 1 and Ax. 2. □

Lemma 2. $D(A) \wedge (A \equiv B) \rightarrow D(B)$.

Proof. By Ax. 2. □

Lemma 3. $D(A) \wedge (A \equiv B) \rightarrow (A \leftrightarrow B)$.

Proof. By Ax. 7 and Lemma 2. Suppose $D(A)$, $(A \equiv B)$, and A. Then $A \equiv \mathbf{t}$, for if $A \equiv \mathbf{f}$ then $\neg A$ by Ax. 7; so $B \equiv \mathbf{t}$, so B by Ax. 7. Thus $A \rightarrow B$; similarly, $B \rightarrow A$. □

Lemma 4. $D(A) \rightarrow (T(A) \leftrightarrow A)$.

Proof. By Ax. 1 and Lemma 3. □

It may then be seen that the truth conditions for $\neg$, $\vee$, and $\forall$ are as usual for definite formulas. Using the methods of Feferman (2008), the axioms of S can be strengthened to having D be strongly compositional in Ax. 6 (i.e. the implications are replaced by equivalences) and still have the system be conservative over PA.

Discussion. Since S is a conservative extension of PA, it does not satisfy the condition (L2). For if S proved the sentence P expressing that all provable sentences of PA are true, it would follow that S and hence PA itself proves the consistency of PA. Also, S is not immune to the "generalization" problem that has been raised for deflationary theories, i.e. the provability of formal versions of statements such as that for any definite proposition p, $p \vee \neg p$ is true. For, that cannot be expressed in $L(S)$ with the use of D as an operator, not a predicate. However, we can consistently extend the axioms DT of Part I (i.e., of Feferman 2008) into the language $L(S)$ by addition of Ax. 1 and some of the other axioms of S, in which such generalizations can be expressed and proved, since there $D(x)$ is written for $T(x) \vee T(\neg.x)$.

5 Conclusion

Most of the preceding has been devoted to the considerations in each example of which of the Leitgeb criteria (L1)-(L8) are to be accepted and which are to be rejected, and little to the reasons for axiomatization given in Sec. 1, though those are always in the background. But given those reasons, I would urge the pursuit of axiomatizations of semantical or definitional approaches that have not yet been thus treated, and the close examination of them in the light of the given criteria.

Acknowledgments

I would like to thank Kentaro Fujimoto and the referee for a number of useful comments on a draft of this paper. Thanks also to Hannes Diener for converting it from a Word file to a LaTeXfile.

References

Aczel, Peter and Solomon Feferman (1980), *Consistency of the unrestricted abstraction principle using an intensional equivalence operator*, in (J. P. Seldin and J. R. Hindley, eds.), To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism, Academic Press, New York, 67-98.

Barwise, Jon and John Etchemendy (1987), *The Liar: An Essay on Truth and Circularity*, Oxford University Press, Oxford.

Blackburn, Simon and Keith Simmons (eds.), (1999), *Truth*, Oxford University Press, Oxford.

Burgess, Alexis G. and John P. Burgess (2011), *Truth*, Princeton University Press, Princeton.

Cantini, Andrea (1996), *Logical Frameworks for Truth and Abstraction: An Axiomatic Study*, vol. 135 of Studies in Logic and the Foundations of Mathematics, Elsevier, Amsterdam.

Feferman, Solomon (1984), *Toward useful type-free theories I*, J. Symbolic Logic 49, 75–111; reprinted in Martin (1984), 237–287.

_____ (1991), *Reflecting on incompleteness*, J. Symbolic Logic 56, 1–49.

_____ (1996), *Gödel's program for new axioms: Why, where, how and what?* in (P. Hajek, ed.) Gödel '96, vol. 6 of Lecture Notes in Logic, 3–22.

_____ (2008), *Axioms for determinateness and truth*, Review Symbolic Logic 1, 204–217.

Feferman, Solomon and Thomas Strahm (2000), *The unfolding of non-finitist arithmetic*, Annals of Pure and Applied Logic 104, 75-96.

_____ (2010), *The unfolding of finitist arithmetic*, Review of Symbolic Logic 3, 665–689.

Field, Hartry (2008), *Saving Truth from Paradox*, Oxford University Press, Oxford.

Friedman, Harvey and Michael Sheard (1987), *An axiomatic approach to self-referential truth*, Annals of Pure and Applied Logic 40, 1–10.

Fujimoto, Kentaro (2010), *Relative truth definability of axiomatic theories*, Bull. Symbolic Logic 16, 305-344.

Gupta, Anil and Nuel Belnap (1993), *The Revision Theory of Truth*, MIT Press, Cambridge MA.

Halbach, Volker (2011), *Axiomatic Theories of Truth*, Cambridge University Press, Cambridge.

Halbach, Volker and Leon Horsten (2006), *Axiomatizing Kripke's theory of truth*, J. Symbolic Logic 71, 677–712.

Horsten, Leon (2011), *The Tarskian Turn: Deflationism and Axiomatic Truth*, MIT Press, Cambridge, MA.

Horwich, Paul (1990), *Truth*, Basil Blackwell, Oxford; 2nd edition, 1998, Clarendon Press, Oxford.

Kirkham, Richard L. (1995), *Theories of Truth: An Introduction*, MIT Press, Cambridge MA.

Kripke, Saul (1975), *Outline of a theory of truth*, J. of Philosophy 72, 690-712; reprinted in Martin (1984), 53–81.

Leitgeb, Hannes (2007), *What theories of truth should be like (but cannot be)*, Blackwell Philosophy Compass 2/2, 276–290.

Lynch, Michael P. (ed.) (2001), *The Nature of Truth: Classical and Contemporary Perspectives*, MIT Press, Cambridge, MA.

Martin, Robert L. (ed.) (1970) *Paradox of the Liar*; reprinted (1978), Ridgeview Pub. Co., Reseda, CA.

_____ (ed.) (1984), *Recent Essays on Truth and the Liar Paradox*, Oxford University Press, New York.

Martin, Robert L. and Peter Woodruff (1975), *On representing "True-in-L" in L*, Philosophia 5, 213-217; reprinted in Martin (1984), 47–51.

Maudlin, Tim (2004), *Truth and Paradox: Solving the Riddles*, Clarendon Press, Oxford.

McDonald, B. E. (2000), On meaningfulness and truth, J. Philosophical Logic 29, 433–482.

McGee, Vann (1991), *Truth, Vagueness and Paradox: An Essay on the Logic of Truth*, Hackett Pub. Co., Indianapolis.

Ramsey, Frank (1927), *Facts and propositions*, Aristotelian Society Supplement 7, 153–170; reprinted in part in Blackburn and Simmons (1999), 106–107.

Reinhardt, William (1985), *Remarks on significance and meaningful applicability*, in (L. P. de Alcantara, ed.) Mathematical Logic and Formal Systems: A Collection of Papers in Honor of Professor Newton C. A. Da Costa, vol. 94 of Lecture Notes in Pure and Applied Mathematics, 227–242.

_____ (1986), Some remarks on extending and interpreting theories, with a partial predicate for truth, J. Philosophical Logic 15, 219–251.

Russell, Bertrand (1908), *Mathematical logic as based on the theory of types*, Amer. J. of Mathematics 30, 222–262; reprinted in (J. van Heijenoort, ed.) From Frege to Gödel: A Source Book in Mathematical Logic (1967), Harvard University Press, Cambridge MA, 150–182.

Sheard, Michael (2002), *Truth, provability, and naïve criteria*, in (V. Halbach and L. Horsten, eds.), Principles of Truth, Dr. Hänsel-Hohenhausen, Frankfurt a. M., 169–181.

Tarski, Alfred (1935), *Der Wahrheitsbegriff in die formalisierten Sprachen*, Studia Philosophica 1, 261–405, English translation: The concept of truth in formalized languages, in Tarski (1956), 152–278.

_____ (1956), Logic, Semantics and Metamathematics: Papers from 1923 to 1938 (translated into English by J. H. Woodger), Clarendon Press, Oxford; 2nd revised edition (1983), (J. Corcoran, ed.) Hackett Pub. Co., Indianapolis.