Volker Halbach • Leon Horsten
Principles of Truth

EPISTEMISCHE STUDIEN
Schriften zur Erkenntnis- und Wissenschaftstheorie

Edited by

Michael Esfeld • Stephan Hartmann • Mike Sandbothe

Band 1 / Volume 1

Volker Halbach • Leon Horsten

Principles of Truth

Second Edition

ontos
verlag
Frankfurt ▪ Lancaster

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie;
detailed bibliographic data is available in the Internet at http://dnb.ddb.de

P.O. Box 15 41 • 63133 Heusenstamm nr. Frankfurt
www.ontosverlag.com

ISBN 3-937202-45-5

2004

Printed on acid-free paper
ISO-Norm 970-6

Printed in Germany.

Preface

The present volume is a record of the lectures given at the conference *Truth, Necessity and Provability*, which was held in Leuven, Belgium, from 18 to 20 November 1999. Except for the paper by Shapiro, all papers included in this volume are based on lectures given at the conference.

On the one hand, the concept of truth is a major research subject in analytical philosophy. On the other hand, mathematical logicians have developed sophisticated logical theories of truth and the paradoxes. The aim of the conference out of which the present volume grew was to bring together prominent logicians and philosophers concerned with truth and the paradoxes. We wanted to promote a deeper interaction and collaboration between them than existed so far. The leading motivation was that recent developments in logical theories of the semantical paradoxes are highly relevant for philosophical research on the notion of truth and that, conversely, philosophical guidance is necessary for the development of logical theories of truth and the paradoxes.

The present volume is therefore intended both for analytical philosophers working on truth and for logicians who concern themselves with the paradoxes. The contributions in this volume present an overview of recent work that has been done on the interface between these two domains.

The amount of knowledge of analytical philosophy and of mathematical logic that is required for understanding the papers in this volume is of a quite modest level. Nevertheless, in order to make the contributions as accessible as possible both to analytical philosophers (graduate students and professional philosophers) and to logicians, the collection of papers is preceded by an extended introductory historical overview of the main results and tenets in philosophical and logical research on truth since the 1930s.

The papers have been grouped into three parts. We have opted for one cumulative bibliography at the end of the book.

Acknowledgements. Ingrid Lombaerts took excellent care of many of the practical aspects concerning the organisation of the conference from which this volume originates. We do not know how we would have managed without her.

We are grateful to all authors we invited for their contributions to this volume and for their patience with this project.

We thank the managing editor Rafael Hüntelmann and the editors of the book series *Epistemische Studien*, Michael Esfeld, Stephan Hartmann and Mike Sandbothe, for their support of this project.

Christopher von Bülow did most of the LaTeX-typesetting and proof-reading. We are indebted to him for the painstaking preparation of the final version and for correcting some mistakes in the introduction. All remaining inconsistencies in the typesetting are due to the editors.

Special thanks go to Ina Ratzke for her help with Word Perfect.

Volker Halbach & Leon Horsten March 2002

Note on the second edition

The second edition of the volume became necessary after only two years. It differs from the first edition only in some small details in the first, introductory chapter. In particular, we corrected a mistake that was brought to our notice by Gabriel Uzquiano. We thank him for his observation.

Volker Halbach & Leon Horsten March 2004

Contents

Volker Halbach and Leon Horsten

Contemporary Methods for Investigating the Concept of Truth An Introduction

Truth as a logico-mathematical concept

Truth is ubiquitous in traditional philosophy as a venerable and deep notion. Most traditional accounts of truth and modern substantialist theories of truth hold that truth is a philosophical notion that needs to be explained in philosophical terms such as correspondence, utility, coherence. It therefore comes as no surprise that truth is exactly the kind of notion that logical empiricism tried to ban from philosophy altogether. Yet today, truth is one of the important research topics in analytic philosophy. Whence the change?

In the early thirties Tarski rehabilitated truth as a respectable notion in his famous work "The Concept of Truth in Formalized Languages" (1935). He gave a definition of truth for a formal language in purely logical and mathematical terms. Tarski's method is fairly general. This suggested that truth is a logico-mathematical notion, and not an irreducible, properly philosophical notion as traditional philosophy would have it. Like all other logico-mathematical concepts, truth is most naturally investigated in formal settings, in the context of other logical and mathematical notions. Attempts should be made to analyze truth in these formal contexts, and the resulting definitions, statements and theorems ought to be regarded as logico-mathematical truths.

Tarski's contemporary heirs are the deflationists. There exists a variety of mutually incompatible theories of truth nowadays that are called deflationary. Deflationism with respect to truth therefore is a somewhat vague or at least underdetermined notion. Moreover, all such accounts differ in several

respects from Tarski's own views on truth. We will see, for example, that most deflationary theories of today incorporate an axiomatization of truth rather than a definition. Nevertheless deflationists side with Tarski in his insistence that truth is a logico-mathematical concept, and the axiomatic approach and Tarski's definitional one are intimately connected, as will become obvious below.

On the formal side, logicians have investigated Tarski's theory and its variants using the methods of proof- and recursion theory. Moreover, Tarski's work on truth in the thirties was the start of modern model theory. Some of these results are relevant also to the deflationary theories. For instance, recently certain proof- and model-theoretic results on conservativeness have proven to be important in the discussion on deflationism.

Most formal work of the last quarter century focused on type-free theories. In these theories the truth predicate can be applied to sentences that contain the same truth predicate themselves.

In the following we browse through some important varieties of logico-mathematical theories. They include Tarski's theory, the relatively vague family of deflationist theories, as well as formal axiomatic and semantical theories of truth.

Tarski's theory of truth

Tarski's theory is the point of departure for most, if not all, recent work on truth. Rather than giving a historically precise account of Tarski's theory we elaborate those features that are most important for later developments. At some points we mention results and tools not known or available to Tarski, but which turn out to be important for Tarski's account.

We choose a setting that is different from Tarski's original framework but which is more suitable for our present concerns. This allows us, for instance, to define truth directly, while Tarski had to take the detour via a satisfaction predicate. Avoiding this detour is only possible in certain special cases.

Tarski's truth predicate pertains to a specific formal language, the so-called object language. We choose the language $\mathcal{L}_{\mathsf{PA}}$ of arithmetic in order to illustrate Tarski's approach. Besides the usual vocabulary of first-order predicate logic with identity, it has a constant symbol 0 for zero, a function symbol S for the successor function, and the binary function symbols $+$ and $\cdot$ for the operations of addition and multiplication, respectively. We do not presuppose that the object language is interpreted in any way or that a deductive system for it is given. We only need to know what the well-formed expressions of $\mathcal{L}_{\mathsf{PA}}$ are.

The truth predicate does not belong to the object language $\mathcal{L}_{\mathsf{PA}}$. It forms part of the metalanguage $\mathcal{L}_{\mathrm{M}}$. In the simplest case the object language is a sublanguage of the metalanguage, i.e., all formulas of the object language are also formulas of the metalanguage. If the metalanguage does not include the object language, then a *translation* of the object- into the metalanguage is required. Whether a translation is correct depends of course on the meaning of the sentences of the object language, whatever "meaning" is. Thus the need for a translation has caused many worries. We stick to the simple case and assume that $\mathcal{L}_{\mathsf{PA}}$ is a sublanguage of the metalanguage $\mathcal{L}_{\mathrm{M}}$.

Tarski develops his theory at first in natural language enriched by some mathematical symbols. However, Tarski is aware of the requirement for a precisely defined metalanguage that allows for strict formal proofs in it.

For the metalanguage $\mathcal{L}_{\mathrm{M}}$ Tarski needs some rules governing the use of the expressions of the object language. Tarski assumed that we have axioms and rules in the language $\mathcal{L}_{\mathrm{M}}$ that allow for derivations within the metalanguage. He called the resulting theory "Metawissenschaft", which is usually translated as "metatheory". Tarski stated some conditions that should be satisfied by the metatheory. It is a deductive system with axioms and rules for $\mathcal{L}_{\mathrm{M}}$. This theory contains the theory of the object language (if there is any) and allows to prove certain facts about expressions.

In order to distinguish the metatheory from the pure uninterpreted and unaxiomatized metalanguage, we denote the former by MT and the latter by $\mathcal{L}_{\mathrm{M}}$.

Tarski assumed that the metalanguage $\mathcal{L}_{\mathrm{M}}$ has a *name* for each sentence of the object language. However, not any kind of name will do. It must be possible to read off the form of the sentence from its name. For this purpose Tarski introduced his *structural-descriptive* names. We do not discuss Tarski's original approach but we present two examples of naming systems that comply with the requirement that one must be able to recover the shape of an expression from its name.

In natural language the *quotational* name of a sentence satisfies the condition that the name has to reveal the structure of the sentence: the singular term

"Snow is white"

designates the sentence within the quotation marks and thus the name displays the exact shape of the sentence.

For mathematical languages, *codings* are used. A coding (or Gödel numbering) associates with any expression e of, say, $\mathcal{L}_{\mathsf{PA}}$ a natural number $\ulcorner e \urcorner$ in an effective way. The details of a typical coding are given, e.g., by Smoryński 1985. For the number $\ulcorner e \urcorner$ we have a name in the language $\mathcal{L}_{\mathsf{PA}}$: for any

number n, the expression

$$\underbrace{S(S(\ldots S}_{n \text{ times}}\underbrace{(0)\ldots))}_{n}$$

is its *numeral* $\overline{n}$. For instance, the numeral $\overline{3}$ of the number 3 is the $\mathcal{L}_{\mathsf{PA}}$-expression $S(S(S(0)))$. Via the coding we obtain names for sentences in the language $\mathcal{L}_{\mathsf{PA}}$: we employ the numeral $\overline{\ulcorner A \urcorner}$ of the code $\ulcorner A \urcorner$ as a name for any sentence A of $\mathcal{L}_{\mathsf{PA}}$.

Since the metalanguage $\mathcal{L}_{\mathrm{M}}$ contains the object language $\mathcal{L}_{\mathsf{PA}}$ and therefore also all numerals, $\mathcal{L}_{\mathrm{M}}$ has names for all sentences of the object language $\mathcal{L}_{\mathsf{PA}}$.

Tarski aimed at a definition of truth. A (potential) definition of truth in the metalanguage $\mathcal{L}_{\mathrm{M}}$ plus the symbol T is a definition of the primitive predicate symbol T in the metalanguage $\mathcal{L}_{\mathrm{M}}$ which takes the following form:

$$\forall x \left(\mathrm{T}x \leftrightarrow \mathit{TDef}(x)\right), \tag{1}$$

where $\mathit{TDef}(x)$ is a complex formula in the metalanguage $\mathcal{L}_{\mathrm{M}}$, containing one free variable (x). The resulting theory is a definitional extension of the theory MT. We are sloppy and call the formula $\mathit{TDef}(x)$ itself a truth definition.

Of course, the sentence (1) is the general pattern for introducing new unary predicates by an explicit definition. Thus, obviously, not just any choice of $\mathit{TDef}(x)$ is acceptable as a definition of *truth*. However, there may be different acceptable choices. What are the distinguishing features of an *adequate* definition of truth?

Tarski's answer to this question is his material adequacy condition. Whether a definition is materially adequate depends on the metatheory MT. The metatheory must prove certain things about the predicate $\mathit{TDef}(x)$ that is supposed to define truth.

In the terms of our present setup involving $\mathcal{L}_{\mathsf{PA}}$ as object language and a metalanguage including $\mathcal{L}_{\mathsf{PA}}$, Tarski's condition can be rephrased as follows:

> A definition of truth is *adequate* if and only if the sentence
>
> $$\mathit{TDef}(\overline{\ulcorner A \urcorner}) \leftrightarrow A$$
>
> is provable in MT for every sentence A of $\mathcal{L}_{\mathsf{PA}}$.

The sentences $\mathit{TDef}(\overline{\ulcorner A \urcorner}) \leftrightarrow A$ are Tarski's famous T-sentences; together they form the T-scheme. Tarski (1935) requires for the adequacy also that

MT proves that only sentences (or their codes) are true; we neglect this additional restriction, which is also not essential according to Tarski.

Whether there is a materially adequate truth definition patently depends on the metatheory. One obvious choice does not work: if MT is a theory in the language $\mathcal{L}_{\mathsf{PA}}$ extending the theory R of arithmetic (Robinson's arithmetic), then it can contain no adequate definition of truth. This observation is a slight variant of Tarski's famous theorem on the undefinability of truth for $\mathcal{L}_{\mathrm{M}}$.

Robinson's arithmetic R is an extremely weak theory. It is much weaker than the better-known theory Peano arithmetic (PA). Thus the theorem on undefinability tells us that the metalanguage $\mathcal{L}_{\mathrm{M}}$ must properly extend the object language $\mathcal{L}_{\mathsf{PA}}$, if there is to be an adequate truth definition. Moreover, the metatheory MT must feature axioms and/or rules for this additional vocabulary that allow to derive all T-sentences.

In order to define truth, Tarski first defined the notion of satisfaction, and from this he defined truth. As pointed out above, we can avoid this detour and define truth directly by induction on the complexity of sentences:

1. If s and t are closed terms of $\mathcal{L}_{\mathsf{PA}}$ with the same value, then $s = t$ is true.
2. If the sentence A is not true, then $\neg A$ is true.
3. If A and B are true, so is $A \wedge B$.
4. If all sentences $A(\overline{n})$ are true, so is $\forall x\, A(x)$.

Here A and B are sentences of $\mathcal{L}_{\mathsf{PA}}$. In the last clause it is assumed that bound variables are renamed in the case of a variable collision.

This recursive definition of truth can be formalized in a sufficiently strong subtheory of second-order arithmetic. The theory ACA (not used by Tarski) that features a comprehension scheme for arithmetical sets and an induction scheme for all second-order formulas suffices for proving the existence of a set of formulas satisfying the above clauses. Thus one can construct a formula $\mathrm{True}(x)$ in the language of second-order arithmetic such that the theory ACA proves the following equivalence for all sentences of $\mathcal{L}_{\mathsf{PA}}$:

$$\mathrm{True}(\overline{\ulcorner A \urcorner}) \leftrightarrow A.$$

Therefore $\mathrm{True}(x)$ (or rather the result of substituting $\mathrm{True}(x)$ for $TDef(x)$ in Formula 1 on the facing page) is a materially adequate definition of truth.

ACA is not the weakest subsystem of second-order arithmetic that allows for a materially adequate definition of truth in Tarski's sense. Tarski was

aware of the fact that different metatheories can provide materially adequate theories. Later logicians like Mostowski (1950) gave a more detailed analysis of what exactly is needed in the metatheory, though the latter worked in a set-theoretic context.

This rough sketch of Tarski's theory and its later developments concludes our exposition of the (type-free) semantical theory of truth. Tarski's paper also contained a few tentative remarks on axiomatic theories.

Tarski tried to dispel the neopositivist distrust of truth by providing a definition of truth, at least for certain restricted areas of application. But in his article "The Concept of Truth in Formalized Languages" he already considered axiomatic approaches, which he finally rejected. For instance, Tarski considered the option of taking the T-sentences as axioms.

One may speculate that he was sceptical of such approaches and opted for a reduction of the concept of truth by definition because it was not fashionable in neopositivism to take truth as a primitive undefined symbol.

Tarski's official arguments against axiomatic approaches, however, are different. According to Tarski, the T-sentences do not prove certain important principles. For instance, they do not prove the universal sentence saying that for any sentence the sentence itself or its negation is true. Furthermore Tarski suggests that any strengthening of the T-sentences will remain incomplete and leave out other important principles (see also Halbach 2000a). He apparently did not suspect that "nice" axiomatizations of truth would be found. This topic recurs today in the discussion of deflationism, as we shall show below.

Deflationism

Unlike Tarski, most deflationists today do not attempt to define truth. Two factors have played a role here. First, Tarski's theory diminished the distrust in the notion of truth. His work was taken to have shown that truth is a safe concept. Second, as time went by, a shift in attitude toward second-order logic gradually took place. Until the thirties, second-order logic was regarded as relatively unproblematic, and even as a part of logic according to logicism. Second-order logic was used without reservations by Russell and Whitehead, for instance. But from the fifties onwards, second-order logic was regarded with mounting suspicion. It became increasingly clear that systems of second-order logic are quite powerful. From a philosophical point of view, Quine insisted that second-order logic is really set theory in disguise, and therefore is not without ontological commitments. Thus second-order logic and axiomatized systems of second-order logic slowly

came to be regarded with apprehension. But if a definition of truth is considered unattainable because of such a definition's entanglement with second-order logic, why not take truth to be a primitive notion?

In the following we will explore the main varieties of axiomatic theories of truth. Fortunately deflationists can build on the work of logicians, in particular proof theorists, who have worked on axiomatic theories of truth. Their aim was not to provide foundations for deflationism. Rather they mostly focused on the role of truth in the foundations of mathematics, where truth turned out to be as useful as certain kinds of second-order quantification. However, their work comes in handy for the deflationist.

Typed axiomatic theories of truth

By *typed* theories we mean deductive systems that do not prove the truth of any sentence containing the truth predicate T. Typed systems cannot prove

$$\mathrm{T}\,\overline{\ulcorner \mathrm{T}\,\overline{\ulcorner \overline{0} = \overline{0} \urcorner}\urcorner},$$

for instance. They are opposed to type-free systems, which are also known as theories of self-referential truth.

We shall start to sketch simple and deductively weak truth theories and then proceed to the stronger and more sophisticated theories (see Sheard 1994 and Halbach 2000b for more extensive overviews). The reader less interested in the technical elaborations may skip this section.

Suppose we want to construct an axiomatic theory of truth for $\mathcal{L}_{\mathsf{PA}}$. By employing the T-sentences as the sole axioms for truth, one obtains a minimal theory of truth, for every materially adequate theory of truth must contain the T-sentences as theorems. Thus a theory of truth based only on the T-sentences is materially adequate and contains no superfluous additional commitments.

Formally, this theory is formulated in the expanded language $\mathcal{L}_{\mathrm{T}}$, which is the language $\mathcal{L}_{\mathsf{PA}}$ of arithmetic plus an additional unary predicate symbol T. The theory is the closure under first-order logic of PA plus the infinite collection of all sentences of the form

$$\mathrm{T}\,\overline{\ulcorner A \urcorner} \leftrightarrow A,$$

with A any sentence of $\mathcal{L}_{\mathsf{PA}}$. Let us call this theory the *Disquotational Theory of Truth* (DT) for $\mathcal{L}_{\mathsf{PA}}$.

There are a couple of remaining issues to be settled here. First, one may wonder whether in instances of the induction scheme of DT, occurrences of

the truth predicate T are allowed. Call the theory which is just like PA, except that T may occur in instances of the induction scheme, PA^{T}. Obviously DT, which is based on PA, is weaker than the corresponding system based on PA^{T}. One may also wonder whether in the biconditionals, A can be allowed to contain free variables. The problem, of course, is how to bind free variables in $\ulcorner A \urcorner$. In arithmetic this can be achieved. For instance, if x is free in A then one says formally that the result of substituting the numeral x for the variable x in A is true. This way, one can quantify into the "Gödel corners".

We indicate that A may contain free variables that can be bound from outside by using the bracket notation "$[A]$". Thus one may look at the universal closure of the following sentences, where A is a formula of $\mathcal{L}_{\mathsf{PA}}$ with arbitrary free variables:

$$\mathrm{T}[A] \leftrightarrow A.$$

Augmenting PA by these stronger axioms yields a theory proving more sentences. But again it does not prove more arithmetical sentences than PA or DT (see Halbach 1999a).

DT and its variants are attractive for deflationists: the axioms are simple, truly deflationary, and they rely on the "disquotational feature" of truth. This fits many deflationary accounts of truth where truth is described as nothing more than a device of disquotation. Furthermore, DT and the variants mentioned do not really have substantial consequences in a sense to be explained below. The deductive weakness, however, is also the Achilles' heel of deflationary accounts of truth based on the T-sentences as the only axioms.

This became clear when opponents of deflationism, e.g., Gupta (1993a), attacked deflationist accounts like Horwich's (1998b) minimalism because they seem to rely on axioms akin to the T-sentences. Nevertheless, it is important to bear in mind that in his contribution to this volume, Horwich emphasizes that his aim was, and is, not to construct a *theory* of truth, in the sense of a set of fundamental theoretical postulates on the basis of which all other facts about truth can be explained. Rather, as he puts it, his approach purports to specify which of the nonsemantic facts about the word 'true' is responsible for its meaning what it does. And his suggestion is that our allegiance to the T-sentences is the key here.

In order to illustrate the deductive weakness of DT, we employ a further notational convention for the bracket notation: if a sentence A occurs only within the brackets "$[\ldots]$" within another formula, then A is treated like a free variable. For instance, in $\mathrm{T}[\neg A] \rightarrow \neg \mathrm{T}[A]$ we may quantify over A. If we prefix a quantifier ranging over (codes of) sentences A of $\mathcal{L}_{\mathsf{PA}}$ then we

obtain the sentence saying that for all sentences A, if $\neg A$ is true then A itself is not true. We refer the reader to the literature for an exact account.

DT is deductively weak because it does not prove certain general principles that may be expected to be provable in a good theory of truth. This observation, as well as the following example, is due to Tarski (1935). DT does not prove the following sentence prefixed by a universal quantifier ranging over all sentences A of $\mathcal{L}_{\mathsf{PA}}$:

$$\mathrm{T}[A] \vee \mathrm{T}[\neg A]; \tag{2}$$

DT only proves the corresponding *scheme*, i. e., DT only proves

$$\mathrm{T}\ulcorner A \urcorner \vee \mathrm{T}\ulcorner \neg A \urcorner$$

for all sentences A of $\mathcal{L}_{\mathsf{PA}}$, but not the universal principle. This deductive limitation holds also for the stronger variant of DT considered above, which is based on PA^{T} and in which the T-sentences take the form of universal closures.

Already Tarski (1935) rejected theories that are based on the T-sentences as the only axioms for truth. If the T-sentences are the only acceptable truth axioms for the deflationist, then, epigrammatically speaking, Tarski refuted deflationism prior to its birth.

If one wants one's truth theory to prove such general principles like the above law of excluded third, then one has to add further axioms. Consequently deflationists tried to overcome this deficiency by opting for stronger theories. For instance, Field (1994a), Tennant (2002) and McGrath (1997) (building on Sosa 1993) have proposed theories that prove more than just the T-sentences or similar principles. These theories overcome the deductive weakness of the truth theories that have been used by deflationists previously.

By Gödel's incompleteness theorems all theories extending DT are incomplete. Thus a sound theory can never be complete in the sense that it decides all sentences formulated in $\mathcal{L}_{\mathrm{T}}$. However, one can still hope to arrive at a complete theory in the sense that the theory proves all truth-theoretic principles we may ever think of. Such a theory would not prove certain sentences, because of the first incompleteness theorem. However, we do not expect a good theory of truth to decide those sentences. For there is no good truth-theoretic axiom for the truth predicate of $\mathcal{L}_{\mathsf{PA}}$ that settles these sentences, or so one might argue.

In fact, the theory $\mathsf{PA}(S)$ which we are going to describe now may have such a status. It does not decide all sentences of $\mathcal{L}_{\mathsf{PA}}$, let alone all sentences of $\mathcal{L}_{\mathrm{T}}$. The examples of such undecided sentences fall into two categories:

either they are Gödel sentences, consistency statements, etc., or they are strong combinatorial principles. We should not expect that these sentences can be decided by invoking a truth predicate for $\mathcal{L}_{\mathsf{PA}}$. All known proofs of them rely on a truth predicate for more inclusive languages like $\mathcal{L}_{\mathrm{T}}$, or on further assumptions like strong second-order axioms.

We now turn to the formulation of the theory $\mathsf{PA}(S)$. Aside from Tarski's material adequacy condition, another core intuition concerning truth is that the truth predicate distributes over the logical connectives and the quantifiers. That is, for instance, a sentence $A \wedge B$ is true if and only if both A and B are true. This intuition lies behind Tarski's inductive definition of truth sketched on page 15.

The theory $\mathsf{PA}(S)$ is obtained by turning the clauses for every connective and quantifier into axioms. For simplicity we assume that $\mathcal{L}_{\mathsf{PA}}$ has only $\neg$, $\wedge$ and $\forall$ as its logical symbols. Like DT, the theory $\mathsf{PA}(S)$ is again a typed theory, that is, it is a theory of truth for the language $\mathcal{L}_{\mathsf{PA}}$.

$\mathsf{PA}(S)$ is formulated in $\mathcal{L}_{\mathrm{T}}$. It is the first-order closure of PA with induction in $\mathcal{L}_{\mathrm{T}}$ plus the axioms obtained by closing the following formulas with quantifiers restricted to sentences (and formulas with at most one free variable, in the last clause) of $\mathcal{L}_{\mathsf{PA}}$:

1. $\mathrm{T}[A] \leftrightarrow A$, if A is an atomic formula of $\mathcal{L}_{\mathsf{PA}}$,
2. $\mathrm{T}[A \wedge B] \leftrightarrow \mathrm{T}[A] \wedge \mathrm{T}[B]$,
3. $\mathrm{T}[\neg A] \leftrightarrow \neg \mathrm{T}[A]$,
4. $\mathrm{T}[\forall x\, A] \leftrightarrow \forall y\, \mathrm{T}\big[A(y/x)\big]$.

The last axiom says that a universally quantified formula is true if and only if all its instances are true. If x occurs in A in the scope of a quantifier $\forall y$, then we rename the variable y before y is finally substituted for x. Of course, $A(y/x)$ then is the result of this substitution.[1]

$\mathsf{PA}(S)$ is also called the *Compositional Theory of Truth*. The axioms reduce the truth of complex sentences to the truth of less complex sentences.

Since the axioms of $\mathsf{PA}(S)$ prove all T-sentences, DT is a subtheory of $\mathsf{PA}(S)$. The theory $\mathsf{PA}(S)$ is much stronger than DT. For instance,

[1] [This note is intended for the technically-inclined reader, and can be skipped by others without loss of continuity.] One might wonder why a new variable y is needed and why one does not employ $\mathrm{T}[\forall x\, A] \leftrightarrow \forall x\, \mathrm{T}[A]$ as an axiom instead. The difference may be seen by looking at nonstandard models. In Axiom 4, x may be a nonstandard variable because it occurs only within the brackets "$[\ldots]$", while y has to be standard. Consequently, Axiom 4 tells us that any sentence $\forall x\, A$ is true (for any, possibly nonstandard, variable x) if and only if all its numerical instances are true. In contrast, the alternative formulation says this only for standard variables.

Axiom 3 proves the principle of excluded third (Formula 2 on page 19), which is not provable in DT.

Conservativeness

Variants of DT, and, to a lesser extent, PA(S), are the theories of truth of choice for deflationists, at least if one sticks to typed theories.

We have already mentioned some obvious proof-theoretic properties of DT, its variants, and PA(S). In the recent discussion on deflationism, further proof-theoretic observations have been used.

Some deflationists even claim that truth is a purely logical device. Deflationists generally do not use mathematical tools for setting up their theory. They axiomatize truth, and thus they do not need mathematics for defining it. But this alone does not establish that their concept of truth is purely logical. If truth actually is purely logical, then it should be "neutral". It should not decide "substantial" nonlogical matters; the theory should not yield consequences underivable by pure logic (or perhaps a little bit more). In technical terms, these philosophers claim that their theories of truth are *conservative*.

Clearly, the weaker the theory, the more likely it is to be conservative. Thus the T-sentences are the deflationist's best bet with respect to conservativeness.

However, the T-sentences themselves already prove "something about the world": they prove that there are at least two different objects, because the codes of a tautology and of a contradiction must be different (see Halbach 2001b). In other words, the T-sentences are nonconservative over the underlying first-order logic. In the light of this trivial observation it is highly implausible to say that truth is a *logical* concept. For even disquotational truth is not ontologically neutral, as one might expect from a logical notion.

The ontological commitment of the T-sentences is very small: they do not prove the existence of further objects, i.e., the T-sentences do not exclude that all true sentences are identical (and the same for all false sentences). Now, even first-order predicate logic is not completely neutral; for it already proves that at least one thing exists.[2] In the light of this, we should not be concerned too much because our theory of truth proves the existence of a further object. This becomes plausible if we agree with Tarski that a theory of truth without an underlying theory of the objects which may

[2]For this reason many philosophers claim that *free logic*, which does not prove the existence of anything, is more adequate as pure logic.

be true does not make much sense. Tarski wanted such a theory to form part of his metatheory. Thus a deflationist could still try to save a kind of conservativeness by claiming that a notion of truth should be conservative over the underlying arithmetical theory, Peano arithmetic.

Somehow, the minimal ontological commitment of the T-sentences has to be tolerated by all deflationists. However, as the contributions to this volume bear out, there is disagreement in the deflationist camp whether an acceptable theory of truth ought to be conservative over the arithmetical theory to which it is added. If it were conservative then truth would be neutral at least in the context of a certain presupposed theory. At least in the context of PA truth would be like a logical notion because it would not allow for new mathematical insights. In this respect truth would be free of mathematical commitments.

The theory DT is successful in this respect: DT is conservative over PA, that is, DT does not prove any sentence that is not already a consequence of PA. As we have seen, however, DT is hardly sufficient as a theory of truth because it is deductively too weak.

Thus we turn to PA(S). This theory proves new arithmetical truths that are not provable in PA, whence it is not arithmetically conservative over PA. In fact PA(S) is proof-theoretically strong: it does not only prove the consistency of PA, but many more and stronger arithmetical truths (see, e.g., Feferman 1991). In this sense, it appears that truth yields new "substantial" insights.

Ketland (1999) and Shapiro (1998) have attacked deflationism on these grounds: How can a theory of truth be "deflationary" if it proves mathematically substantial theorems? The commitments of the theory PA(S) are the same as those of the second-order theory ACA mentioned earlier (see again Feferman 1991). Hence the truth predicate of PA(S) is surely not deflationary with respect to its mathematical consequences.

In response, Field (1999) maintains that deflationism implies only that PA(S) without the induction axioms involving truth has to be conservative over PA. However, it must be kept in mind here that although this theory is indeed conservative over PA, this fact is far from obvious. Kotlarski, Krajewski, and Lachlan (1981) employed an ingenious model-theoretic argument in order to prove that PA(S) with arithmetical induction only is conservative over PA.[3] That the proof cannot be trivial follows from a deep theorem

[3]In the first edition we erroneously wrote that PA(S) is conservative over PA, where we should have written that PA(S) with induction restricted to the arithmetical language is conservative. The mistake was pointed out by Gabriel Uzquiano in his review in the *Notre Dame Philosophical Reviews* published at http://ndpr.icaap.org/content/archives/2003/4/uzquiano-halbach.html.

proved by Lachlan (1981), according to which only very special models of PA can be expanded to models of PA(S). Recently, proof-theoretic arguments for the conservativeness of PA(S) (with arithmetical induction only) over PA have also become available (Halbach 1999a).

Some deflationists, however, explicitly distance themselves from the commitment that an acceptable theory of truth must be arithmetically conservative. It seems that in this context, the distinction between considering truth to be a *logical* notion and considering truth to be a logico-*mathematical* notion really makes a difference.

Even the truth predicate of PA(S) is deflationary in the sense that it is mathematical. I.e., the theory is independent of the physical world and does not depend on any contingent features.

Semantic theories of type-free truth

Tarski's inductive definition of truth is a semantic theory. The axiomatic counterpart is the theory PA(S) because PA(S) has the inductive clauses as axioms. The recursion-theoretic properties of the semantic construction and the proof-theoretic properties of the corresponding typed theories are fairly well understood.

Logicians sought new challenges by studying type-free theories of truth. Many different theories have been developed and in this introduction we cannot even attempt to sketch all of the most important ones. The reader may turn to the monographs Brendel 1992, McGee 1991, Gupta and Belnap 1993, Cantini 1996, and Halbach 1996.

Here, we restrict ourselves to a particularly successful semantical theory and its axiomatizations, namely Kripke's theory of truth. Several logicians experimented with partial logic, supervaluations and transfinite inductive definitions in order to set up type-free theories of truth. Kripke (1975) took up these developments, unified them in a very general setting, and found phenomena that had gone unnoticed. Moreover, he availed himself of the theory of positive inductive definitions that had been studied in detail and in generality shortly before by Moschovakis (1974) and others.

Kripke exploited the fact that Tarski's theorem on the undefinability of truth does not apply in an unqualified manner to partially interpreted languages. He constructed partial models for $\mathcal{L}_{\mathsf{PA}}$ of which it can be said in a somewhat weakened sense that they make the unrestricted Tarski-biconditional sentences true. His models are formed in stages, indexed by ordinals. The simplest of Kripke's models, the so-called least-fixed-point model of the strong Kleene scheme, is constructed along the following lines.

One wants to build a model for the language $\mathcal{L}_{\mathrm{T}}$. The arithmetical vocabulary is interpreted throughout as in the standard model $\mathbb{N}$; the truth predicate T will be the only partially interpreted symbol: it will receive, at each ordinal stage, an extension $\mathcal{E}$ and an anti-extension $\mathcal{A}$. The union $\mathcal{E} \cup \mathcal{A}$ does not exhaust the domain; otherwise T would be a total predicate. The extension $\mathcal{E}$ of T is the collection of (codes of) sentences which are (at the given stage) determinately true; the anti-extension $\mathcal{A}$ of T is the collection of (codes of) sentences which are (at the given stage) determinately false. A partial model $\mathcal{M}$ for $\mathcal{L}_{\mathrm{T}}$ can then be identified with the ordered pair $(\mathcal{E}, \mathcal{A})$. In general, the union of the extension and the anti-extension will not exhaust the collection of all sentences: some sentences will at each ordinal stage retain their indeterminate status. An example of an eternally indeterminate sentence is the Liar sentence L such that

$$\mathsf{PA} \vdash L \leftrightarrow \neg \mathrm{T} \overline{\ulcorner L \urcorner}.$$

The intuition that the Liar argument tells us that the Liar sentence L cannot have a determinate truth-value is the basic motivation for constructing a theory of truth in which T is treated as a partial predicate.

At stage 0, one lets the extension $\mathcal{E}_0$ and the anti-extension $\mathcal{A}_0$ be empty. This yields a partial model, $\mathcal{M}_0 = (\mathcal{E}_0, \mathcal{A}_0) = (\emptyset, \emptyset)$. Next, one uses a popular evaluation scheme for partial logic, the so-called *strong Kleene scheme*. The strong Kleene evaluation scheme $\vDash_{\mathrm{sk}}$ is defined as follows:

1. For any atomic formula $Fx_1 \ldots x_n$: $\mathcal{M} \vDash_{\mathrm{sk}} Fx_1 \ldots x_n$ (resp., $\mathcal{M} \vDash_{\mathrm{sk}} \neg Fx_1 \ldots x_n$) iff the n-tuple $(o_1, \ldots, o_n)$, where o_i is assigned to x_i for $i = 1, \ldots, n$, belongs to the extension (anti-extension) of F.

2. For all formulas A, B:

   - $\mathcal{M} \vDash_{\mathrm{sk}} A \wedge B$ iff $\mathcal{M} \vDash_{\mathrm{sk}} A$ and $\mathcal{M} \vDash_{\mathrm{sk}} B$;
   - $\mathcal{M} \vDash_{\mathrm{sk}} \neg(A \wedge B)$ iff either $\mathcal{M} \vDash_{\mathrm{sk}} \neg A$ or $\mathcal{M} \vDash_{\mathrm{sk}} \neg B$ (or both);
   - $\mathcal{M} \vDash_{\mathrm{sk}} \forall x\, A$ iff for all n, $\mathcal{M} \vDash_{\mathrm{sk}} A(\overline{n}/x)$;
   - $\mathcal{M} \vDash_{\mathrm{sk}} \neg\forall x\, A$ iff for at least one n, $\mathcal{M} \vDash_{\mathrm{sk}} \neg A(\overline{n}/x)$.

One now uses this strong Kleene scheme to determine the collection $\mathcal{E}_1$ of sentences of $\mathcal{L}_{\mathrm{T}}$ that are made true by $\mathcal{M}_0$, and the collection $\mathcal{A}_1$ of sentences of $\mathcal{L}_{\mathrm{T}}$ the negation of which is made true by $\mathcal{M}_0$. Thus a new partial model $\mathcal{M}_1 = (\mathcal{E}_1, \mathcal{A}_1)$ is obtained. Using $\mathcal{M}_1$, a new extension $\mathcal{E}_2$ and a new anti-extension $\mathcal{A}_2$ are then determined on the basis of $\mathcal{M}_1$, and

so on. In general, for any ordinal α,

$$\mathcal{E}_{\alpha+1} \equiv \{ A \in \mathcal{L}_{\mathrm{T}} \mid \mathcal{M}_\alpha \vDash_{\mathrm{sk}} A \}$$

and

$$\mathcal{A}_{\alpha+1} \equiv \{ A \in \mathcal{L}_{\mathrm{T}} \mid \mathcal{M}_\alpha \vDash_{\mathrm{sk}} \neg A \}.$$

For limit stages λ, we set

$$\mathcal{E}_\lambda \equiv \bigcup_{\kappa<\lambda} \mathcal{E}_\kappa$$

and

$$\mathcal{A}_\lambda \equiv \bigcup_{\kappa<\lambda} \mathcal{A}_\kappa.$$

This process eventually (after transfinitely many stages) comes to a stage ∞ where no new sentences enter the extension of T or the anti-extension of T. The partial model $\mathcal{M}_\infty = (\mathcal{E}_\infty, \mathcal{A}_\infty)$ is called the *least fixed point* of the strong Kleene hierarchy of partial models. For every sentence A of $\mathcal{L}_{\mathrm{T}}$,

- A is true in $\mathcal{M}_\infty$ under the strong Kleene scheme if and only if $\mathrm{T}\ulcorner\overline{A}\urcorner$ is true in $\mathcal{M}_\infty$;
- $\neg A$ is true in $\mathcal{M}_\infty$ if and only if $\neg\mathrm{T}\ulcorner\overline{A}\urcorner$ is true in $\mathcal{M}_\infty$;
- A is undetermined by $\mathcal{M}_\infty$ if and only if $\mathrm{T}\ulcorner\overline{A}\urcorner$ is undetermined by $\mathcal{M}_\infty$.

In this precise sense, $\mathcal{M}_\infty$ makes a weak form of the T-sentences true. The Liar sentence L remains undetermined in $\mathcal{M}_\infty$, as it should be, on the underlying motivation of Kripke's approach.

There are alternatives to using the strong Kleene scheme in Kripke's construction. Kripke himself emphasized that one could also use van Fraassen's *supervaluation scheme*. The supervaluation scheme is defined as follows. For all partial models $\mathcal{M}$, $\mathcal{N}$, say that $\mathcal{M} \subseteq \mathcal{N}$ ("$\mathcal{N}$ extends $\mathcal{M}$") if for every predicate F of the language, the extension (anti-extension) of F according to $\mathcal{M}$ is a subset of the extension (anti-extension) of F according to $\mathcal{N}$. Then the supervaluation evaluation scheme $\vDash_{\mathrm{sv}}$ is obtained by changing the inductive clause in the definition of the strong Kleene scheme into

2′. For any formulas A, B:

- $\mathcal{M} \vDash_{\mathrm{sv}} A \wedge B$ iff for every total extension $\mathcal{N}$ of $\mathcal{M}$: $\mathcal{N} \vDash A$ and $\mathcal{N} \vDash B$;