Logic and **L**ogical **P**hilosophy
Volume 15 (2006), 285–303
DOI: 10.12775/LLP.2006.017

Roman Murawski

TROUBLES WITH (THE CONCEPT OF) TRUTH IN MATHEMATICS*

Abstract. In the paper the problem of definability and undefinability of the concept of satisfaction and truth is considered. Connections between satisfaction and truth on the one hand and consistency of certain systems of ω-logic and transfinite induction on the other are indicated.

Keywords: truth, satisfaction, satisfaction class, nonstandard model, Peano arithmetic, second-order arithmetic, ω-logic, transfinite induction.

Several concepts of truth and several approaches to this concept have been proposed in the logic: coherence theory, correspondence theory, pragmatist theory, redundancy theory and semantic theory. The last one due to Tarski is probably the most influential and most widely accepted theory of truth — though not free of critiques. Tarski hoped that his definition will "catch hold of the actual meaning of an old notion" (Tarski 1944). Since according to him the "old" notion of truth is ambiguous and even doubtfully coherent, he restricted his concern to what he called the "classical Aristotelian conception of truth" as expressed in Aristotle's dictum:

> To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true.

*The financial support of the Committee for Scientific Researches (grant no 1 H01A 042 27) is acknowledged.

Received November 29, 2006
 ISSN: 1425-3305

Tarski's theory falls into two parts: he provided, first, adequacy conditions, i.e. conditions which any acceptable definition of truth ought to fulfil; and then a definition of truth for a specified formal language.

The question of the philosophical significance of Tarski's theory of truth is a hard one. It has been criticized both for saying too little and for saying too much. For example Black wrote in (1948, p. 260):

> the neutrality of Tarski's definition with respect to the competing philosophical theories of truth is sufficient to demonstrate its lack of philosophical relevance.

On the other hand Mackie (1973, p. 40) said that

> The Tarskian theory [...] belongs to factual rather than conceptual analysis [...]. Tarski's theory has plenty of meat to it, whereas a correct conceptual analysis of truth has very little.

Tarski himself was modest about the epistemological pretensions of his theory. Though he was convinced that his concept of satisfaction and truth is a contribution to the philosophical problem of truth (cf. his famous paper 1933), on the other hand he emphasized that his conception is philosophically neutral. In (1944) he wrote:

> we may accept the semantic conception of truth without giving up any epistemological attitude we may have had, we may remain naive realists or idealists, empiricists or metaphysicians. [...] The semantic conception is completely neutral toward all these issues.

Despite of these controversies it is the fact that just Tarski's theory of truth has been accepted in the foundations of mathematics. Hence we shall not discuss the philosophical problems connected with it but we shall indicate some other problems — of a metamathematical and foundational character.

* * *

Tarski provided in (1933) a definition (in a non-formalized metasystem) of satisfaction and truth and, on the other hand, proved a theorem on the undefinability of the concept of truth for a formalized language L in L itself. It was stated as Theorem I (β) and said that:[1]

[1] Cf. Tarski, 1965, p. 247.

> Assuming that the class of all provable sentences of the metatheory is consistent, it is impossible to construct an adequate definition of truth in the sense of convention **T** on the basis of the metatheory.

It was followed by a description of the idea of the proof and then by a sketch of the proof. Note that the theorem was proved by diagonalization.

To fix our attention and to be more precise let us restrict ourselves to Peano arithmetic. This is a first-order theory formalized in the language L(PA) with the following nonlogical symbols: 0, S, $+, \cdot$ and based on the following nonlogical axioms:

(A1) $S(x) = S(y) \to x = y$,

(A2) $\neg(0 = S(x))$,

(A3) $x + 0 = x$,

(A4) $x + S(y) = S(x + y)$,

(A5) $x \cdot 0 = 0$,

(A6) $x \cdot S(y) = x \cdot y + x$,

(A7) $\varphi(0) \wedge \forall x[\varphi(x) \to \varphi(S(x))] \longrightarrow \forall x \varphi(x)$,
where φ is any formula of the language L(PA).

Fix an arithmetization of the language L(PA) and denote by $\ulcorner\varphi\urcorner$ the Gödel number of a formula φ by the given arithmetization.[2] Let $\overline{n}$ be the term $\underbrace{S \ldots S}_{n}(0)$ denoting the natural number n.

The strong version of Tarski's theorem (i.e., the version without parameters) can be now formulated in the following way.

THEOREM 1 (Tarski, 1933). *If Peano arithmetic* PA *is consistent then there exists no formula* $\mathbf{St}(x)$ *of the language* L(PA) *being the definition of truth for formulas of* L(PA)*, i.e., such a formula* $\mathbf{St}(x)$ *that for any sentence* ψ *of the language* L(PA)

$$\mathrm{PA} \vdash \psi \equiv \mathbf{St}(\overline{\ulcorner\psi\urcorner}).$$

Let $\mathcal{N}_0$ be the standard interpretation of the language of Peano arithmetic, i.e., $\mathcal{N}_0 = \langle \mathbb{N}, 0, S, +, \cdot \rangle$ where $\mathbb{N}$ is the set of natural numbers, 0 is the number zero, S is the successor function and $+$ and $\cdot$ are addition and multiplication of natural numbers, resp. The structure $\mathcal{N}_0$ is called the standard

[2] Detailed information on the arithmetization and on the arithmetical counterparts of various metamathematical notions can be found, e.g., in Mendelson (1964), Shoenfield (1967) or Murawski (1999b).

model of PA. Tarski's theorem states that there exists no formula **St** of the language L(PA) such that for any sentence ψ of L(PA), $\mathrm{PA} \vdash \psi \equiv \mathbf{St}(\overline{\ulcorner \psi \urcorner})$, hence in particular there exists no formula **St** such that for any sentence ψ of L(PA), $\mathcal{N}_0 \models \psi$ if and only if $\mathcal{N}_0 \models \mathbf{St}(\overline{\ulcorner \psi \urcorner})$, i.e., there is no definition (in the language of L(PA)) of the set of (Gödel numbers of) those sentences of L(PA) which are true in the domain of natural numbers (= in the standard model $\mathcal{N}_0$). Consequently the notion of truth for arithmetic of natural numbers, i.e., the set

$$\{\ulcorner \varphi \urcorner : \varphi \text{ is a sentence of L(PA) } \& \ \mathcal{N}_0 \models \varphi\}$$

is not an arithmetical set. This contrasts with the fact that the notion of provability for arithmetic, i.e., the set

$$\{\ulcorner \varphi \urcorner : \varphi \text{ is a sentence of L(PA) } \& \ \mathrm{PA} \vdash \varphi\}$$

is an arithmetical set, in fact it is recursively enumerable. This indicates the gap between provability and truth. On the other hand one can show that the notion of truth for arithmetic is hyperarithmetical, i.e., it belongs to the class Δ^1_1.[3]

Tarski's theorem can be easily generalized to theories extending Peano arithmetic PA. In fact the following theorem holds.

THEOREM 2. *Let* T *be any consistent first-order theory extending Peano arithmetic* PA *and let* $\mathcal{M}$ *be any model of* T. *Then the set* $Th(\mathcal{M}) = \{\ulcorner \psi \urcorner : \mathcal{M} \models \psi\}$, *i.e., the set of Gödel numbers of all sentences true in* $\mathcal{M}$, *is not definable in* $\mathcal{M}$.

Note also that in the above theorems only the notion of truth, i.e., of satisfaction of sentences, was considered. One can generalize them of course to the case of satisfaction of formulas with free variables. Let T be an extension of PA (the language L(T) can also be an extension of the language L(PA)).

DEFINITION 3. *A binary predicate* **S** *of the language of* L(T) *is said to be a satisfaction predicate for the theory* PA *in the sense* (A) *if and only if for for every formula* φ *of* L(PA) *all free variables of which occur among variables* $x_1, \ldots, x_n$ *and any natural numbers* $k_1, \ldots, k_n$*:*

$$\mathrm{T} \vdash \varphi(\overline{k_1}, \ldots, \overline{k_n}) \equiv \mathbf{S}(\overline{\ulcorner \varphi \urcorner}, \overline{\langle k_1, \ldots, k_n \rangle}).$$

[3] Cf. Mostowski (1951). For information on the hyperarithmetical hierarchy see Rogers (1967) or Shoenfield (1967).

DEFINITION 4. *A binary predicate* **S** *of the language of* L(T) *is said to be a satisfaction predicate for the theory* PA *in the sense* (B) *if and only if for for every formula* φ *of* L(PA) *all free variables of which occur among variables* $x_1, \ldots, x_n$*:*

$$\mathrm{T} \vdash \forall x\{\mathbf{Seq}(x) \wedge \mathbf{lh}(x) = \overline{n} \rightarrow [\varphi((x)_1, \ldots, (x)_n) \equiv \mathbf{S}(\overline{\ulcorner \varphi \urcorner}, x)]\}.$$

DEFINITION 5. *A binary predicate* **S** *of the language of* L(T) *is said to be a satisfaction predicate for the theory* PA *in the sense* (C) *if and only if the following formulas are provable in* T:

$$\mathbf{S}(u, v) \rightarrow \mathbf{Form}(u) \wedge \mathbf{Seq}(v) \wedge \mathbf{lh}(v) = \mathbf{F}(u),$$

$$\mathbf{Term}(t_1) \wedge \mathbf{Term}(t_2) \wedge u = \overline{\langle SN(=), t_1, t_2 \rangle} \rightarrow \\ \rightarrow [\mathbf{S}(u, v) \equiv \mathbf{val}(t_1, v|\mathbf{F}(t_1)) = \mathbf{val}(t_2, v|\mathbf{F}(t_2))].$$

$$u = \overline{\langle SN(\neg), u_1 \rangle} \wedge \mathbf{Form}(u_1) \rightarrow [\mathbf{S}(u, v) \equiv \neg \mathbf{S}(u_1, v)],$$

$$u = \overline{\langle SN(\vee), u_1, u_2 \rangle} \wedge \mathbf{Form}(u_1) \wedge \mathbf{Form}(u_2) \rightarrow \\ \rightarrow [\mathbf{S}(u, v) \equiv \mathbf{S}(u_1, v|\mathbf{F}(u_1)) \vee \mathbf{S}(u_2, v|\mathbf{F}(u_2))],$$

$$u = \overline{\langle SN(\exists), \ulcorner x_k \urcorner, u_1 \rangle} \wedge \mathbf{Form}(u_1) \wedge \neg \mathbf{Fr}(u_1, 2k) \rightarrow \\ \rightarrow [\mathbf{S}(u, v) \equiv \mathbf{S}(u_1, v)],$$

$$u = \overline{\langle SN(\exists), \ulcorner x_k \urcorner, u_1 \rangle} \wedge \mathbf{Form}(u_1) \wedge \mathbf{Fr}(u_1, 2k) \rightarrow \\ \rightarrow [\mathbf{S}(u, v) \equiv \exists x \mathbf{S}(u_1, v * \binom{k}{x})]$$

where $v * \binom{k}{n}$ *denotes a sequence number* w *such that*

$$lh(w) = \max(lh(v), k),$$
$$\forall i < lh(v)[i \neq k \rightarrow (w)_i = (v)_i],$$
$$(w)_k = x,$$
$$\forall i[lh(v) < i < k \rightarrow (w)_i = 0].$$

Add that we adopt here the following convention: if R is a recursive relation then by $\mathbf{R}$ we denote a formula of the language L(PA) strongly representing R in PA.

Note that Tarski considered in (1933) the notion of a satisfaction predicate in sense (A). Observe also that if $\mathbf{S}$ is a satisfaction predicate in the sense (C) then it is a satisfaction predicate in the sense (B) (this follows by induction) and if $\mathbf{S}$ is a satisfaction predicate in the sense (B) then it is a satisfaction predicate in the sense (A) (this is obvious from the definitions). So Tarski's theorem on undefinability of truth implies that there is no satisfaction predicate for PA in the sense (A) definable in PA. Hence there are no satisfaction predicates in the sense (B) or (C) for PA definable in PA.

A connection between the notion of satisfaction and the notion of consistency is indicated by the following theorem.[4]

THEOREM 6. *Let* T *be an extension of Peano arithmetic* PA *such that induction (with respect to all formulas of the language* L(T)*) holds in* T*. If* $\mathbf{S}$ *is a satisfaction predicate for* PA *in the theory* T *in the sense* (C) *then* T *proves the consistency of* PA*, i.e.,* $\mathrm{T} \vdash \mathrm{Con}_{\mathrm{PA}}$*, where* $\mathrm{Con}_{\mathrm{PA}}$ *denotes the formula* $\neg\mathbf{Pr}(\overline{\ulcorner 0 = \overline{1}\urcorner})$.

Note that the last theorem does not hold for $\mathbf{S}$ being a satisfaction predicate in the sense (A) or (B).

As mentioned above Tarski used in his undefinability theorem Gödel's method of diagonalization. From a historico-philosophical point of view it should be noted that Tarski made clear his indebtedness to Gödel's methods but on the other hand he strongly emphasized the fact that his results had been obtained independently. Gödel was aware of the formal undefinability of the notion of truth in 1931. In fact it was precisely his recognition of the contrast between the formal definability of provability and the formal undefinability of truth that led him to his discovery of incompleteness. Gödel did not mention the undefinability of truth in his writings, he even avoided the terms "truth" and "true", because he feared that work assuming such a concept would be rejected by foundational establishment dominated by Hilbert's ideas. Tarski was free of such limitations. In fact, in the Lvov-Warsaw School no restrictive initial preconditions were assumed before the proper investigations could start. Note also that Gödel had no precise definition of the concept of truth.[5]

[4] The proof of this theorem can be found, e.g., in Murawski's (1999b).

[5] More information on this problem can be found in (Woleński, 1991) and (Murawski,

Having shown that the notion of truth for Peano arithmetic PA cannot be defined in PA itself one should ask *where* it can be defined. We have here two possibilities: (1) one can consider an appropriate extension of PA (possibly weak) in which the notion can be defined and (2) one can extend the language L(PA) by adding a new binary predicate **S** (called satisfaction class) and characterizing it axiomatically by adding to Peano arithmetic PA (as new axioms) sentences given above in the definition of a satisfaction predicate in the sense (C). Note that since those axioms form a finite set of axioms one can write them as a single formula of the language $\mathrm{L(PA)} \cup \mathbf{S}$ (denote it as "**S** is a satisfaction class"). Let us consider both those possibilities.

A natural extension of PA which can be considered in our context is the so-called second-order arithmetic A_2^-. This is a first-order (!) system formalized in a language with two sorts of variables: number variables x, y, z, ... and set variables X, Y, Z, ... Its nonlogical constants are those of Peano arithmetic, i.e., 0, S, $+$, $\cdot$ as well as symbols for all primitive recursive functions and the membership relation $\in$. Nonlogical axioms of A_2^- are the following:

(1) axioms of PA without the axiom scheme of induction,

(2) (extensionality) $\forall x(x \in X \equiv x \in Y) \to X = Y$,

(3) (induction axiom)

$$0 \in X \wedge \forall x(x \in X \to Sx \in X) \to \forall x(x \in X),$$

(4) recursive definitional equations for primitive recursive functions,

(5) (axiom scheme of comprehension)

$$\exists X \forall x[x \in X \equiv \varphi(x, \ldots)],$$

where φ is any formula of the language of A_2^- (possibly with free number- or set-variables) in which X does not occur free.

If Γ is a class of formulas of the language $\mathrm{L}(\mathrm{A}_2^-)$ then we denote by $\mathrm{A}_2^-|\Gamma$ the subsystem of A_2^- obtained by restricting the comprehension axiom to formulas belonging to the class Γ. Later we shall consider in particular the system $\mathrm{A}_2^-|\Sigma_1^1$ where Σ_1^1 is the class of formulas of the form $\exists X \varphi(X, \ldots)$ where φ is an arithmetical formula, i.e., a formula containing possibly any

1998).

quantifiers bounding number-variables and no quantifier over set-variables. One can prove the following theorems:[6]

THEOREM 7. *Second order arithmetic* A_2^- *proves the existence of the satisfaction predicate in the sense* (C) *for Peano arithmetic. Moreover, this can be proved in the fragment* $\mathrm{A}_2^-|\Sigma_1^1$ *of* A_2^-.

Using this theorem and Theorem 6 we obtain

THEOREM 8. $\mathrm{A}_2^-|\Sigma_1^1 \vdash \mathrm{Con}_{\mathrm{PA}}$.

In this way we showed that the notion of truth (in fact the notion of satisfaction in the sense (C)) for L(PA) can be defined in the theory $\mathrm{A}_2^-|\Sigma_1^1$.

It turns out that, in contrast with Tarski's theorem, the notion of satisfaction and truth for certain fragments of the language L(PA) can be defined in Peano arithmetic itself. To formulate precisely appropriate results a hierarchy of formulas of the language L(PA) similar to the arithmetical hierarchy of relations is needed. Let $\Sigma_0^0 = \Pi_0^0 = \Delta_0^0$ be the smallest class of formulas of the language L(PA) containing atomic formulas and closed under connectives and bounded quantifiers. We define Σ_{n+1}^0 to be the set of all formulas equivalent (in PA) to formulas of the form $\exists x\psi$ for $\psi \in \Pi_n^0$ and Π_{n+1}^0 to be the set of all formulas equivalent (in PA) to formulas of the form $\forall x\psi$ for $\psi \in \Sigma_n^0$. We put also Δ_n^0 to be the set of all formulas equivalent (in PA) to a Σ_n^0 formula and to a Π_n^0 formula.

One can show that there exist formulas $Sat_{\Delta_0^0}$, $Sat_{\Sigma_n^0}$ and $Sat_{\Pi_n^0}$ of L(PA) which are definitions of satisfaction for, resp., Δ_0^0, Σ_n^0 and Π_n^0 formulas ($n \in \mathbb{N}$). Moreover, the formula $Sat_{\Delta_0^0}$ can be written as both Σ_1^0 and Π_1^0 formula. Hence one can say that there exists a Δ_1^0 definition of satisfaction for Δ_0^0 formulas of L(PA). Consequently the formulas $Sat_{\Sigma_n^0}$ and $Sat_{\Pi_n^0}$ are, resp., Σ_n^0 and Π_n^0 definitions of satisfaction for Σ_n^0 and Π_n^0 formulas of L(PA) ($n \in \mathbb{N}$). One can also show that the appropriate properties of those formulas (corresponding to the metamathematical properties of the appropriate notions of satisfaction) can be proved in Peano arithmetic PA.[7] Let further $Tr_{\Sigma_n^0}$ and $Tr_{\Pi_n^0}$ denote truth predicates for Σ_n^0 and Π_n^0 sentences.[8] In the sequel we shall identify formulas defining satisfaction and truth and their extensions in the standard model $\mathcal{N}_0$.

[6] Proofs of those theorems can be found in (Murawski, 1999a and 1999b).

[7] In fact it can be proved even in the fragment of Peano arithmetic with induction for Σ_1^0 formulas only. Cf. Kaye (1991), Murawski (1999b) and Hájek-Pudlák (1993).

[8] Construction of $Sat_{\Sigma_n^0}$ and $Sat_{\Pi_n^0}$ can be found in Kaye (1991) and Murawski (1999b).

It turns out that the (partial) truth (in the standard model $\mathcal{N}_0$), i.e., truth for Σ^0_n formulas can be approximated by iterations of the so-called ω-rule. Hence one can say that the (infinitary) ω-rule enables us to express, to reach the partial truth. To be more precise let us introduce the following hierarchies. Let T be any first-order theory in the language L(PA) of Peano arithmetic. The first hierarchy is defined as follows:

$$\begin{aligned}
\mathrm{T}^0 &= \mathrm{T},\\
\mathrm{T}^{\alpha+\frac{1}{2}} &= \mathrm{T}^\alpha \cup \{\varphi : \varphi \text{ is of the form } \forall x\psi(x) \text{ and } \psi(\overline{n}) \in \mathrm{T}^\alpha,\\
&\qquad \text{for every } n \in \mathbb{N}\},\\
\mathrm{T}^{\alpha+1} &= \text{the smallest set of formulas containing } \mathrm{T}^{\alpha+\frac{1}{2}}\\
&\qquad \text{and closed under the rules of inference of PA},\\
\mathrm{T}^\lambda &= \bigcup_{\alpha<\lambda} \mathrm{T}^\alpha \text{ for } \lambda \text{ limit}.
\end{aligned}$$

The second hierarchy is defined so (cf. Niebergall, 1996):

$$\begin{aligned}
\mathrm{T}^{(0)} &= \mathrm{T},\\
\mathrm{T}^{(\alpha+\frac{1}{2})} &= \mathrm{T}^{(\alpha)} \cup \{\varphi : \varphi \text{ is of the form } \forall x\psi(x) \text{ and } \psi(x) \in \Sigma^0_{2\alpha+1}\\
&\qquad \text{and } \psi(\overline{n}) \in \mathrm{T}^{(\alpha)} \text{ for every } n \in \mathbb{N}\},\\
\mathrm{T}^{(\alpha+1)} &= \text{the smallest set of formulas containing } \mathrm{T}^{(\alpha+\frac{1}{2})}\\
&\qquad \text{and closed under the rules of inference of PA},\\
\mathrm{T}^{(\lambda)} &= \bigcup_{\alpha<\lambda} \mathrm{T}^{(\alpha)} \text{ for } \lambda \text{ limit}.
\end{aligned}$$

Hence the ω-rule is now applied at stage n to Σ^0_{2n+1} formulas only.

The last hierarchy is the following one (cf. Niebergall, 1996):

$$\begin{aligned}
(\Sigma^k\mathrm{T})^0 &= \mathrm{T},\\
(\Sigma^k\mathrm{T})^{\alpha+\frac{1}{2}} &= (\Sigma^k\mathrm{T})^\alpha \cup \{\varphi : \varphi \text{ is of the form } \forall x\psi(x) \text{ and } \psi(x) \in \Sigma^0_k\\
&\qquad \text{and } \psi(\overline{n}) \in (\Sigma^k\mathrm{T})^\alpha \text{ for every } n \in \mathbb{N}\},\\
(\Sigma^k\mathrm{T})^{\alpha+1} &= \text{the smallest set of formulas containing } (\Sigma^k\mathrm{T})^{\alpha+\frac{1}{2}}\\
&\qquad \text{and closed under the rules of inference of PA},\\
(\Sigma^k\mathrm{T})^\lambda &= \bigcup_{\alpha<\lambda} (\Sigma^k\mathrm{T})^\alpha \text{ for } \lambda \text{ limit}.
\end{aligned}$$

We still need one notion—reflection principle. So let T be a theory whose set of (Gödel numbers of) theorems is strongly representable in PA. Denote by $RFN(\mathrm{T})$ the uniform reflection principle for T, i.e., the scheme

$$\forall x\ Pr_{\mathrm{T}}(\overline{\ulcorner \varphi(x) \urcorner}) \to \forall x \varphi(x)$$

for $\varphi(x)$ formula of L(T) with at most one free variable. If one restricts the class of formulas to a class Γ (for example Σ^0_k or Π^0_k) then one obtains $RFN_\Gamma(\mathrm{T})$.

The local reflection principle for T denoted by $Rfn(\mathrm{T})$ is the following scheme

$$Pr_{\mathrm{T}}(\overline{\ulcorner \varphi \urcorner}) \to \varphi$$

for φ closed.

We have now the following facts:[9]

THEOREM 9. (1) $\mathrm{PA}^n \supseteq \mathrm{PA} + Tr_{\Sigma^0_{2n+1}}$, *i.e., for every* $n \in \mathbb{N}$ *the theory* PA^n *is complete with respect to* Σ^0_{2n+1} *sentences.*

(2) (Niebergall, 1996) *For any* $n \in \mathbb{N}$, $\mathrm{PA}^{(n)} = \mathrm{PA} + Tr_{\Sigma^0_{2n+1}}$.

(3) (Niebergall, 1996) *For any* $n \in \mathbb{N}$, $(\Sigma^k\mathrm{PA})^{n+1} = \mathrm{PA} + Tr_{\Sigma^0_{k+2}}$, *if* $k \leq 2n$.

(4) (Niebergall, 1996) *For any* $n \in \mathbb{N}$, $\mathrm{PA}^{n+1} = \mathrm{PA} + Tr_{\Sigma^0_{2n+3}} + RFN(\mathrm{PA}^n)$.

(5) (Niebergall, 1996) *For any* $n \in \mathbb{N}$, $(\mathrm{PA} + Tr_{\Sigma^0_k})^n = \mathrm{PA}^n + Tr_{\Sigma^0_{k+2n}}$.

(6) (Feferman, 1962) *For a suitable class of ordinals:* (a) *iterating* $\mathrm{T} \to \mathrm{T} + \mathrm{Con}_{\mathrm{T}}$ *or* $\mathrm{T} \to \mathrm{T} + Rfn(\mathrm{T})$ *one has* $\bigcup \mathrm{PA}_\alpha = \mathrm{PA} + Tr_{\Pi^0_1}$; (b) *iterating* $\mathrm{T} \to \mathrm{T} + RFN(\mathrm{T})$ *one obtains* $\bigcup \mathrm{PA}_\alpha = Th(\mathcal{N}_0) =$ *all true sentences of arithmetic.*

Turn now to the second possibility indicated above, i.e., to the axiomatic characterization of satisfaction and truth. Recall that one extends now the language L(PA) by adding a new binary predicate **S** (called a satisfaction class; denote the new language by L_{S}) and characterizing it axiomatically by adding to Peano arithmetic PA (as new axioms) sentences given above in the definition of a satisfaction predicate in the sense (C). Note that since those axioms form a finite set of axioms one can write them as a single formula of

[9] They are only examples of theorems that should indicate the character of results.

the language L(PA)∪**S** (denote it as "**S** is a satisfaction class"). One can add certain additional axioms stating that **S** has special properties. Two such properties are significant: being full and being inductive. A satisfaction class **S** is said to be full if and only if it decides every formula on any valuation. And **S** is said to be inductive if and only if the induction principle holds for all formulas of the extended language L_S. If Γ is a class of formulas of L_S and one requires that the induction principle holds for all formulas of Γ only then **S** is called Γ-inductive. Denote by Γ − PA(**S**) the theory PA + "**S** is a full Γ-inductive satisfaction class" and by PA(**S**) the theory PA + "**S** is a full inductive satisfaction class".

There arises a question whether theories of the type Γ − PA(**S**) or the theory PA(**S**) are consistent, i.e., whether they have models. Note that if $\langle \mathcal{M}, S\rangle$ is a model of such an extension of PA then $\mathcal{M}$ is a model of PA and S is a satisfaction predicate for L(PA) over the model $\mathcal{M}$ (S is called a satisfaction class over the model $\mathcal{M}$). It turns out that not over every model $\mathcal{M}$ of PA one can define a predicate S such that the structure $\langle \mathcal{M}, S\rangle$ is a model of PA + "**S** is a satisfaction class", i.e., not for every model of PA the notion of satisfaction (truth) (satisfying the natural Tarski's conditions) exists. The crucial property of a model $\mathcal{M}$ needed here is recursive saturation defined as follows:

Definition 10. *A model $\mathcal{M} \models$ PA is said to be recursively saturated iff for every recursive type Θ over the model $\mathcal{M}$, if Θ is consistent over $\mathcal{M}$ then Θ is realized in $\mathcal{M}$.*

In fact the following theorem holds:

Theorem 11. *For any countable model $\mathcal{M}$ of PA the following conditions are equivalent:*

(a) *$\mathcal{M}$ is recursively saturated,*

(b) *$\mathcal{M}$ has a satisfaction class,*

(c) *$\mathcal{M}$ has a full satisfaction class,*

(d) *$\mathcal{M}$ has an inductive satisfaction class.*

From this theorem it follows also that the theories:

PA + "**S** is a satisfaction class",

PA + "**S** is a full satisfaction class",

PA + "**S** is an inductive satisfaction class"

are all conservative extensions of PA, i.e., one can prove in those theories exactly the same theorems about natural numbers (i.e., formulas of the language L(PA)) as in Peano arithmetic PA. Hence the addition of a new notion, i.e., of a notion of a satisfaction (truth), with properties indicated above does not increase the prooftheoretical power of a theory with respect to sentences of the language L(PA). On the other hand the assumption that a satisfaction class is full and Δ_0^0-inductive gives a nonconservative extension of PA! In fact one can prove in this theory, i.e., in Δ_0^0–PA($\mathbf{S}$) the consistency of PA.

This leads us to the problem when does there exist a model of a theory of the type $\Gamma - \mathrm{PA}(\mathbf{S})$ for Γ such that $\Delta_0^0 \subseteq \Gamma$, i.e., when for a model $\mathcal{M}$ of PA does there exist a full Γ-inductive satisfaction class over $\mathcal{M}$? The answer is: the model $\mathcal{M}$ must be recursively saturated and must satisfy certain extension of Peano arithmetic. Those extensions can be characterized in the language of consistency of appropriate ω-logics or of appropriate transfinite induction.

Consider the following sequence of formulas of the language L(PA) (one uses here arithmetization):

$$\Gamma_0(\varphi) = \text{“PA} \vdash \varphi\text{“},$$

$$\Gamma_{n+\frac{1}{2}}(\varphi) = \text{“}\varphi \text{ is of the form } \eta \vee \forall z \psi(z) \text{ and } \forall z \Gamma_n(\eta \vee \psi(S^z 0))\text{”},$$

$$\Gamma_{n+1}(\varphi) = \text{“there exists a proof of } \varphi \text{ based on PA} \cup \{\psi : \Gamma_{n+\frac{1}{2}}(\psi)\}\text{”}.$$

Observe that in this system of ω-logic only the application of the ω-rule increases the degree of complexity of a proof.

THEOREM 12 (Kotlarski, 1986). *Let $\mathcal{M}$ be a countable recursively saturated model of* PA. *Then there exists a full Δ_0^0-inductive satisfaction class over $\mathcal{M}$ iff for any $n \in \mathbb{N}$: $\mathcal{M} \models \neg\Gamma_n(0 = 1)$.*

It can also be proved (cf. Kotlarski, 1986) that the theory $\Delta_0^0 - \mathrm{PA}(\mathbf{S})$ is equal to the theory

$$\mathrm{PA} + \mathbf{S} \text{ is a full satisfaction class} + \forall\varphi[(\mathrm{PA} \vdash \varphi) \to \mathbf{S}(\varphi)].$$

The last sentence can be read as: “$\mathbf{S}$ makes all theorems of PA true”. It is equivalent to the Δ_0^0-inductiveness of the satisfaction class $\mathbf{S}$.

The system of ω-logic described above can be iterated in the transfinite. So let us fix a “natural” system of notations for ordinals $< \varepsilon_0$ (one gets it by Cantor’s Normal Form Theorem). By transfinite induction on $\alpha < \varepsilon_0$ we define theories T^α and formulas Γ_n^α in the following way:

$$\mathrm{T}^0 = \mathrm{PA},$$

$$\Gamma^0_0(\varphi) = \text{“PA} \vdash \varphi\text{”},$$
$$\Gamma^\alpha_0(\varphi) = \text{“}T^\alpha \vdash \varphi\text{”},$$
$$\Gamma^\alpha_{n+\frac{1}{2}}(\varphi) = \text{“}\varphi \text{ is of the form } \eta \vee \forall z\, \psi(z) \text{ and } \forall z\, \Gamma^\alpha_n(\eta \vee \psi(z))\text{”},$$
$$\Gamma^\alpha_{n+1}(\varphi) = \text{“T}^\alpha \cup \Gamma^\alpha_{n+\frac{1}{2}} \vdash \varphi\text{”},$$
$$\mathrm{T}^{\alpha+1} = \mathrm{T}^\alpha \cup \{\neg\Gamma^\alpha_n(0=1) : n \in \mathbb{N}\},$$
$$\mathrm{T}^\lambda = \bigcup_{\alpha<\lambda} \mathrm{T}^\alpha,\ \lambda \text{ limit}.$$

Using Recursion Theorem one can formalize those definitions in PA. Define now for an ordinal α a sequence $\omega_m(\alpha)$ in the following way: $\omega_0(\alpha) = \alpha, \omega_{m+1}(\alpha) = \omega^{\omega_m(\alpha)}$. The following theorem holds.

THEOREM 13 (Kotlarski and Ratajczyk, 1990a). (1) *Let m be a natural number and let $\mathcal{M} \models$ PA be countable and recursively saturated. Then there exists a full Σ^0_m-inductive satisfaction class over $\mathcal{M}$ iff for every $k \in \mathbb{N}$: $\mathcal{M} \models \neg\Gamma^{\omega_m(k)}_k(0=1)$.*

(2) *Let $\mathcal{M}$ be a countable and recursively saturated model of* PA. *Then there exists a full inductive satisfaction class over $\mathcal{M}$ iff for every $n \in \mathbb{N}$:*

$$\mathcal{M} \models \neg\Gamma^{\omega_n}_n(0=1)$$

where $\omega_n = \omega_n(\omega)$.

Let now $TI(\rho)$, where ρ is an ordinal, denote the scheme of transfinite induction up to ρ. Then the following theorem holds.

THEOREM 14 (Kotlarski and Ratajczyk, 1990b). *Let $\mathcal{M}$ be a countable and recursively saturated model of* PA *and let m be a natural number. Then*

(1) *there exists a full Σ^0_m-inductive satisfaction class over the model $\mathcal{M}$ iff for every $k \in \mathbb{N}$, $\mathcal{M}$ satisfies the transfinite induction up to $\varepsilon_{\omega_m(k)}$, i.e., $\mathcal{M} \models TI(\varepsilon_{\omega_m(k)})$,*

(2) *there exists a full inductive satisfaction class over the model $\mathcal{M}$ iff for every $k \in \mathbb{N}$, $\mathcal{M}$ satisfies the transfinite induction up to ε_{ω_k}, i.e., $\mathcal{M} \models TI(\varepsilon_{\omega_k})$.*

The above theorems show that not always a full Γ-inductive satisfaction class does exist. In fact a given model of PA must satisfy additional conditions. Those conditions indicate connections between satisfaction (truth) on

the one hand and transfinite induction and consistency of certain ω-logics on the other.

They do this also in another way. Let T be an extension of Peano arithmetic PA. Define a theory PA^{T} in the following way:

$$\mathrm{PA}^{\mathrm{T}} = \{\varphi \in \mathrm{L(PA)} : \mathrm{T} \vdash \varphi\}.$$

Hence theorems of PA^{T} are those sentences of the language L(PA) of Peano arithmetic (hence sentences about natural numbers) which can be proved in the stronger theory T.

Let now T be a theory of the type of PA(S) or its fragment. How do theories PA^{T} look like? The answer is provided by the following theorem.

THEOREM 15. (i) (Kotlarski, 1986) $\mathrm{PA}^{\Delta_0^0-\mathrm{PA}(S)} = \mathrm{PA} \cup \{\neg\Gamma_n(0 = 1) : n \in \mathbb{N}\}$.

(ii) (Kotlarski and Ratajczyk, 1990a) *Let m be a natural number. Then*

$$\mathrm{PA}^{\Sigma_m^0-\mathrm{PA}(S)} = \mathrm{PA} \cup \{\neg\Gamma_k^{\omega_m(k)}(0 = 1) : k \in \mathbb{N}\},$$
$$\mathrm{PA}^{\mathrm{PA}(S)} = \mathrm{PA} \cup \{\neg\Gamma_n^{\omega_n}(0 = 1) : n \in \mathbb{N}\},$$

where $\omega_n = \omega_n(\omega)$.

(iii) (Kotlarski and Ratajczyk, 1990b) *Let m be a natural number. Then*

$$\mathrm{PA}^{\Sigma_m^0-\mathrm{PA}(S)} = \mathrm{PA} \cup \{TI\,(\varepsilon_{\omega_m(k)}) : k \in \mathbb{N}\},$$
$$\mathrm{PA}^{\mathrm{PA}(S)} = \mathrm{PA} \cup \{TI\,(\varepsilon_{\omega_k}) : k \in \mathbb{N}\}.$$

This theorem shows that what can be proved about natural numbers using Peano axioms and the notion of satisfaction (truth) that is assumed to be full and Σ_m^0-inductive is exactly the same as what can be proved in PA plus transfinite induction for ordinals $\varepsilon_{\omega_m(k)}$ (for all $k \in \mathbb{N}$) or in PA plus appropriate consistency statements. Similarly for PA plus full inductive satisfaction (truth) on the one hand and PA plus transfinite induction for ordinals ε_{ω_k} (for all $k \in \mathbb{N}$) or PA plus appropriate consistency statements on the other. It shows also that by adding to PA the notion of a satisfaction (truth) and assuming that it is full and makes all theorems of PA true one obtains a theory with exactly the same theorems about natural numbers as by taking PA augmented with a concept of a full and Δ_0^0-inductive satisfaction (truth) or PA plus appropriate consistency statements. So (the usage of) satisfaction (truth) can be in a certain sense approximated by transfinite

induction or by adding certain consistency statements concerning appropriate systems of ω-logic. Recall also that if T is PA + "**S** is a full satisfaction class" or PA + "**S** is an inductive satisfaction class" then $\mathrm{PA}^{\mathrm{T}} = \mathrm{PA}$. Hence only the assumption that satisfaction (truth) is inductive and full gives new information about natural numbers.

Next problem is the problem of uniqueness: so assume that over a given model $\mathcal{M}$ there exists a satisfaction class. Is it determined uniquely, i.e., does $\mathcal{M}$ admit exactly one satisfaction class or do there exist a variety of them over $\mathcal{M}$? In other words: if a theory in the language $\mathrm{L_S}$ extending PA has a model then can it posses also other models with the same fixed part corresponding to the language L(PA)? The answer is given by the following theorems.

Theorem 16 (Krajewski, 1976). *For any countable model $\mathcal{M}$ of* PA *which admits a full satisfaction class S there exists a countable model $\mathcal{M}_1$ such that* (1) $\mathcal{M}_1 \equiv \mathcal{M}$ *and* (2) *there exist $2^{\aleph_0}$ full satisfaction classes over the model $\mathcal{M}_1$ which are mutually inconsistent on sentences and $(\mathcal{M}, S) \equiv (\mathcal{M}_1, S_\alpha)$ for $\alpha < 2^{\aleph_0}$.*

Theorem 17 (Kossak, 1985). *If there exists a full inductive satisfaction class over a countable model $\mathcal{M}$ then*

(1) *there exist $2^{\aleph_0}$ full inductive satisfaction classes over $\mathcal{M}$ which are pairwise elementarily inequivalent, i.e., such that*

$$(\mathcal{M}, S_{\alpha_1}) \not\equiv (\mathcal{M}, S_{\alpha_2})$$

for $\alpha_1 < \alpha_2 < 2^{\aleph_0}$,

(2) *there exist $2^{\aleph_0}$ full inductive satisfaction classes over $\mathcal{M}$ which are elementarily equivalent but pairwise nonisomorphic.*

Explain that two satisfaction classes over a given model $\mathcal{M}$ are said to be mutually inconsistent on sentences iff there exists a (nonstandard) sentence φ (i.e., a formula without free variables) of the language Form($\mathcal{M}$) such that $S_1(\varphi, \emptyset)$ and $S_2(\neg\varphi, \emptyset)$ or *vice versa*. Hence one of satisfaction classes over $\mathcal{M}$ (i.e., one of the notions of satisfaction for the language Form($\mathcal{M}$)) says that the sentence φ is true and the other says that φ is false! In general, satisfaction classes S_1 and S_2 over $\mathcal{M}$ are said to be mutually inconsistent iff there exists a formula φ in the sense of the model $\mathcal{M}$ and an M-valuation a for the formula φ such that $S_1(\varphi, a)$ and $S_2(\neg\varphi, a)$ or *vice versa*. Hence one

of satisfaction classes says that the formula φ is satisfied on the valuation a and the other one says that φ is not satisfied on a!

Add that two models are said to be elementarily equivalent if and only if they satisfy exactly the same sentences (i.e., formulas without free variables).

The above theorems show that the axiomatic characterization of satisfaction and truth is non-unique. The reason is that Tarski's conditions put on satisfaction classes are too weak and do not uniquely determine the satisfaction and truth. What more, they admit various interpretations, even mutually inconsistent on sentences! Hence the classical principle of bivalency is not any longer valued for nonstandard languages. Moreover, one can find mutually inconsistent satisfaction classes being elementarily equivalent, i.e., having the same elementary properties in the language L(PA) with predicate **S**.

* * *

Let us turn to conclusions. As Gaifman (2004, p. 15) wrote:

> Intended interpretations are closely related to realistic conceptions of mathematical theories. By subscribing to the standard model of natural numbers, we are committing ourselves to the objective truth or falsity of number-theoretic statements, where these are usually taken as statements of first-order arithmetic. The standard model is supposed to provide truth-values for these statements.

Deductive systems can only yield recursively enumerable sets of theorems and therefore they can only partially capture truth in the standard model. Even more, the truth in the standard model is not arithmetically definable.

On the other hand there are nonstandard (hence unintended) models (not only for Peano arithmetic but even for the theory of the standard model $\mathcal{N}_0$). This shows an essential shortcoming of a formalized approach: the failure to fully determine the intended model.

An attempt to define arithmetical truth (truth for arithmetic) in a higher order theory, for example in the second-order arithmetic or its appropriate fragment where its existence can be proved, does not give a satisfactory solution. Indeed second-order arithmetic as a deductive system is incomplete and, additionally, there appears the problem of nonstandard models and interpretations.

So we are forced to attempt to characterize the concept of truth (for PA or for other theories) in an axiomatic way. But here again we encounter

the phenomenon of nonstandardness. In fact, considering a nonstandard[10] model $\langle \mathcal{M}, S\rangle$ for the theory $\Gamma - \mathrm{PA}(\mathbf{S})$ or its fragment we have that $\mathcal{M}$ is a nonstandard model of PA and S is the appropriate satisfaction class over $\mathcal{M}$, hence the satisfaction class for formulas of the language $\mathrm{Form}(\mathcal{M})$ consisting of all those elements of the universe M (standard and nonstandard numbers) that (from the point of view of $\mathcal{M}$) are (i.e., behave like) formulas (identified here with their Gödel numbers). Among them there are also nonstandard formulas, i.e., objects that formally behave like formulas but have no proper metamathematical meaning (they are formulas from the point of view of the world of $\mathcal{M}$, but not from the point of view of the real metamathematical world). Of course $\mathrm{L(PA)} \subseteq \mathrm{Form}(\mathcal{M})$ and

$$S_{tr} = \{(\ulcorner\varphi\urcorner, a) : \varphi \text{ standard formula of L(PA) } a \text{ M-valuation for } \varphi, \mathcal{M} \models \varphi[a]\} \subseteq S.$$

But this "real" satisfaction S_{tr} (and consequently also "real" truth) cannot be arithmetically defined in ("cut" from) the satisfaction class S. Indeed, the notion of being standard is not arithmetically definable.

Theories of the type $\Gamma - \mathrm{PA}(\mathbf{S})$ have a rich variety of models. But on the other hand not every model $\mathcal{M}$ of PA can be extended to a model $\langle \mathcal{M}, S\rangle$ of $\Gamma - \mathrm{PA}(\mathbf{S})$—indeed, the structure $\mathcal{M}$ must satisfy appropriate conditions that can be characterized in the language of consistency of certain systems of ω-logic or of the transfinite induction. This shows also that the usage of satisfaction (truth) in proving theorems about natural numbers (i.e., proving properties of natural numbers in theories of the type $\mathrm{PA}^{\Gamma-\mathrm{PA}(S)}$) can be in a certain sense approximated by transfinite induction or by adding certain consistency statements concerning appropriate systems of ω-logic.

Moreover, even for a fixed model $\mathcal{M}$ of Peano arithmetic for which there exists a satisfaction class, the concept of satisfaction and truth cannot be uniquely determined and, even worse, not always can be defined in such a way that the required (and expected because useful) nice metamathematical properties would be satisfied. There is no uniqueness and no bivalency (for nonstandard models). But nonstandard models and nonstandard languages (generated by such models and by axiomatic approach to the concept of truth) turn out to be useful and to have an impressive spectrum of applications. In particular they can be used to establish properties of deductive

[10] It is impossible to exclude nonstandard models and to restrict ourselves to the standard one only since the latter cannot be characterized arithmetically (in an axiomatic way).

systems, provide insight into fragments of Peano arithmetic as well as into (second-order) expansions of it. They can also serve as a heuristic guide for behavior of the infinity (one can code by nonstandard objects appropriate infinite sets, in particular infinite sets of standard formulas).

Note also that considering satisfaction classes and truth for the language of Peano arithmetic and attempting to characterize them axiomatically we use the whole time at the metatheoretical level Tarski's definition with respect to structures of the type $\langle \mathcal{M}, S \rangle$ and the latter is understood as being defined in a non-formalized metasystem.

A general moral of our considerations is that semantics needs infinitistic means and methods. Hence finitistic tools and means proposed by Hilbert in his programme are essentially in sufficient.

References

Black, M., 1948, "The semantic definition of truth", *Analysis* 8.

Feferman S., 1962, "Transfinite recursive progressions of axiomatic theories", *Journal of Symbolic Logic* 27, 259–316.

Gaifman H., 2004, "Non-standard models in a broader perspective", in: A. Enayat and R. Kossak (eds.), *Nonstandard Models of Arithmetic and Set Theory*, Contemporary Mathematics vol. 361, American Mathematical Society, Providence, Rhode Island, 1–22.

Hájek P. and P. Pudlák, 1993, *Metamathematics of First-Order Arithmetic*, Springer-Verlag, Berlin–Heidelberg–New York.

Kaye R., 1991, *Models of Peano Arithmetic*, Clarendon Press, Oxford.

Kotlarski, H., 1986, "Bounded induction and satisfaction classes", *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 32, 531–544.

Kotlarski H. and Z. Ratajczyk, 1990a, "Inductive full satisfaction classes", *Annals of Pure and Applied Logic* 47, 199–223.

Kotlarski H. and Z. Ratajczyk, 1990b, "More on induction in the language with a full satisfaction class", *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 36, 441–454.

Mackie J.L., 1973, *Truth, Probability and Paradox*, Oxford University Press, Oxford.

Mendelson E., 1964, *Introduction to Mathematical Logic*, D. Van Nostrand Company, Inc., Princeton-Toronto-New York-London.

Mostowski A., 1951, "A Classification of logical systems", *Studia Philosophica* 4, 237–274.

Murawski R., 1998, "Undefinability of truth. The problem of priority: Tarski vs Gödel", *History and Philosophy of Logic* 19, 153–160.

Murawski R., 1999a, "Undefinability vs. definability of satisfaction and truth", in: J. Woleński and E. Köhler (eds.), *Alfred Tarski and the Vienna Circle*, Kluwer Academic Publishers, Dordrecht–Boston–London, 203–215.

Murawski R., 1999b, *Recursive Functions and Metamathematics*, Kluwer Academic Publishers, Dordrecht–Boston–London.

Niebergall, K.-G., 1996, *Zur Metamathematik nichtaxiomatisierbarer Theorien*, Centrum für Informations- und Sprachverarbeitung Ludwig-Maximilians-Universität München, Bericht 96–87.

Rogers H., Jr., 1967, *Theory of Recursive Functions and Effective Computability*, Mc-Graw Hill, New York–St. Luis–San Francisco–Toronto–London–Sydney.

Shoenfield J.R., 1967, *Mathematical Logic*, Addison-Wesley, Reading, Mass.

Tarski A., 1933, *Pojęcie prawdy w językach nauk dedukcyjnych* (The Notion of Truth in Languages of Deductive Sciences), Nakładem Towarzystwa Naukowego Warszawskiego, Warszawa.

Tarski A., 1944, "The semantic conception of truth", *Philosophy and Phenomenological Research* 4, 341–375.

Tarski A., 1965, "The concept of truth in formalized languages", in: *Logic, Semantics, Metamathematics. Papers From 1923 To 1938*, Clarendon Press, Oxford, pp. 152–278.

Woleński, J., 1991, "Gödel, Tarski and the undefinability of truth", in: *Yearbook 1991 of the Kurt Gödel Society* (Jahrbuch 1991 der Kurt-Gödel-Gesellschaft), Wien, pp. 97–108. Reprinted in: J. Woleński, *Essays in the History of Logic and Logical Philosophy*, Jagiellonian University Press, Kraków 1999, pp. 134–138.

Roman Murawski
Adam Mickiewicz University
Faculty of Mathematics and Comp. Sci.
ul. Umultowska 87
61–614 Poznań, Poland
rmur@amu.edu.pl