THE REVIEW OF SYMBOLIC LOGIC
Volume 10, Number 1, March 2017

MODELS OF PT^{-} WITH INTERNAL INDUCTION FOR TOTAL FORMULAE

CEZARY CIEŚLIŃSKI, MATEUSZ ŁEŁYK, and BARTOSZ WCISŁO
Institute of Philosophy, University of Warsaw

Abstract. We show that a typed compositional theory of positive truth with internal induction for total formulae (denoted by PT_{tot}) is not semantically conservative over Peano arithmetic. In addition, we observe that the class of models of PA expandable to models of PT_{tot} contains every recursively saturated model of arithmetic. Our results point to a gap in the philosophical project of describing the use of the truth predicate in model-theoretic contexts.

§1. Introduction. For quite a while, conservativity has been promoted as an important trait of deflationary truth theories. To our knowledge, the idea was introduced for the first time by Leon Horsten, who declared conservativity to be a commitment of Horwich's 'minimal theory'.[1] Since then Horsten's proposal of explaining the deflationary tenet of 'thinness' or 'neutrality' of truth in terms of conservativity has been often repeated, discussed, and refined in logical and philosophical literature.

At least two notions of conservativity have been proposed as tools for the deflationist to explicate his position. One notion is syntactic; the intended meaning is that a conservative extension does not prove new theorems of the base language. The second is semantic and concerns the possibility of expanding models.[2]

DEFINITION 1.1. *Let T_1 and T_2 be theories in languages L_1 and L_2 (with $L_1 \subseteq L_2$). Then,*

(i) *T_2 is syntactically conservative over T_1 iff $\forall \psi \in L_1\ [T_2 \vdash \psi \rightarrow T_1 \vdash \psi]$.*

(ii) *T_2 is semantically conservative over T_1 iff every model of T_1 can be expanded to a model of T_2.*

Although semantic conservativity entails conservativity in the syntactic sense, these two notions are not equivalent. There are well-known examples of axiomatic truth theories which are syntactically, but not semantically conservative over their base theories of syntax.

Research on axiomatic theories of truth permits us to recognise the following fairly general patterns, exemplified by many truth theories.

- Axiomatic truth theories which are both compositional and fully inductive are not syntactically conservative over their base theories of syntax. However, the lack of even one of these properties typically produces syntactic conservativity.[3]

Received: April 2, 2016.

[1] In Horsten's own words: 'The minimalist theory entails that a truth predicate should be conservative over a given theory that is stated without the truth predicate (or any other semantical notions)'. See Horsten (1995), p. 183.

[2] See Cieśliński (2015a) for a discussion of the philosophical motivation behind both versions of the conservativity demand for deflationary theories of truth.

[3] Compositional theories CT, KF, and FS are syntactically nonconservative over PA, but they become conservative as soon as the extended induction—that is, induction for formulae

doi:10.1017/S1755020316000356

- Even without compositionality, full induction in the language with the truth predicate typically produces semantic nonconservativity of an axiomatic theory of truth. However, it is known that compositionality by itself (that is, without extended induction) can be squared with semantic conservativity.[4]

The initial reception of these and related results has been mainly negative. Some authors (see in particular Shapiro (1998) and Ketland (1999)) employed the conservativity demand as a weapon against deflationary theories of truth, arguing that deflationary truth *should* be conservative, but immediately adding that *it cannot be*, because conservative truth theories are too weak. However, more recently some defences of conservative truth theories—even in a stronger semantic sense of the word—have been put forward in the literature. In particular, Fischer & Horsten (2015) proposed to study axiomatic truth theories treated as characterisations of the use of the truth predicate in model-theoretic contexts. In their own words,

> There are contexts where one is reluctant to privilege one model over another, and where one does not want a theory of truth to exclude models for the original language. In particular, this is the case in the single mathematical field where truth predicates play a major role, viz. *model theory*, and in uses of model theory in proof theory. (Fischer & Horsten, 2015, p. 345)

In effect, Fischer and Horsten describe their endeavour as similar in important respects to that of Tarski, who attempted to establish 'beyond reasonable doubt that the uses of truth predicates in metamathematics are legitimate'. (Fischer & Horsten, 2015, p. 345)

Apart from semantic conservativity, there is also another desirable trait of the aforementioned 'uses of truth predicates'. Truth has expressive power—it 'widens the class of thoughts that we can express' (Fischer & Horsten, 2015, p. 345). One indication of the expressive power is the noninterpretability of our theory of truth in its base theory of syntax; there is then a precise sense in which truth brings conceptually something new. Another indication is the nonelementary speed-up of the theory of truth with respect to its arithmetical base theory. Here the expressive power of truth manifests itself in the instrumental value of the truth predicate, namely, truth permits us to shorten proofs. All in all, the moral is that we should search for an axiomatic theory of truth Th which jointly satisfies the following requirements:

(a) Th is semantically conservative over its arithmetical base theory of syntax B,

(b) Th is not interpretable in B,

(c) Th has nonelementary speed-up over B.

Do we have at our disposal an axiomatic truth theory which would capture the uses of the truth predicate in model theory? It has been suggested that a theory of typed positive

with the truth predicate—is removed (see Halbach (2011) for further details). On the other hand, disquotational noncompositional axiomatic theories TB, UTB, and PTB are syntactically conservative, even though they contain full extended induction (see Cieśliński, 2015b and Cieśliński, 2011). A curious counterexample is a disquotational theory PUTB discussed in Halbach (2009), which is not syntactically conservative over Peano arithmetic.

[4] If a theory of truth proves biconditionals '$T(\varphi) \equiv \varphi$' for all arithmetical sentences φ and in addition contains full extended induction, then it is not semantically conservative over PA (see Cieśliński, 2015b). On the other hand, compositional positive truth axioms without any extended induction produce theories which are semantically conservative over PA (see Halbach (2011) for the details).

truth with internal induction for total formulae (denoted as $\mathrm{PT_{tot}}$ in this paper[5]) fills the bill. Below we introduce the relevant definitions.

Let L_{PA} be the language of Peano arithmetic, with Var and Tm^c being (respectively) the sets of variables and constant arithmetical terms. By L_T we denote the extension of L_{PA} with the new one-place predicate '$T(x)$'. For the sentences of L_{PA}, the notation '$Sent_{L_{PA}}$' will be used. By $\underline{x}$ we mean the x-th numeral, i.e., the only numeral denoting the number x. We define the following:

DEFINITION 1.2. PT$^-$ *is the theory in the language L_T which is axiomatised by the usual axioms of Peano arithmetic (*PA*), together with the following truth theoretic axioms*:

(1) $\forall s \forall t \in Tm^c\big(T(s = t) \equiv val(s) = val(t)\big)$,
(2) $\forall s \forall t \in Tm^c\big(T(\neg s = t) \equiv val(s) \neq val(t)\big)$,
(3) $\forall \psi \in Sent_{L_{PA}}\big(T(\neg\neg\psi) \equiv T(\psi)\big)$,
(4) $\forall \varphi \forall \psi \in Sent_{L_{PA}}\big(T(\varphi \wedge \psi) \equiv T(\varphi) \wedge T(\psi)\big)$,
(5) $\forall \varphi \forall \psi \in Sent_{L_{PA}}\big(T\neg(\varphi \wedge \psi) \equiv T(\neg\varphi) \vee T(\neg\psi)\big)$,
(6) $\forall v \in Var \forall \varphi \in L_{PA}\big(T(\forall v \varphi) \equiv \forall x T(\varphi(\underline{x}/v))\big)$,
(7) $\forall v \in Var \forall \varphi \in L_{PA}\big(T(\neg\forall v \varphi) \equiv \exists x T(\neg\varphi(\underline{x}/v))\big)$.

The above axiomatisation is the same as presented in Fischer (2009). In Fischer & Horsten (2015) the considered theory has been augmented with two new axioms to the effect that truth is extensional and only sentences are true, i.e., with the following *extensionality principle*

$$\forall \phi \in L_{PA} \forall s, t \in Tm^c\Big(val(t) = val(s) \rightarrow \big(T(\phi(t)) \equiv T(\phi(s))\big)\Big) \tag{EXT}$$

and the *normality principle*

$$\forall x\big(T(x) \rightarrow x \in Sent_{L_{PA}}\big). \tag{NORM}$$

Although at some point we will consider also these additional axioms, we emphasise that in the terminology adopted in this paper they do not belong to PT$^-$ proper.

In the next move, we extend PT$^-$ with a weak form of induction for total arithmetical formulae.

DEFINITION 1.3. *Let* $\mathrm{tot}(\phi)$ *be a shorthand for* $\forall x\ [T(\phi(x)) \vee T(\neg\phi(x))]$.[6] *By the principle of internal induction for total formulae* (Ind_{tot}), *we mean the following single sentence of the language* L_T:

$$\forall \phi(x) \in L_{PA}\ \Big(\mathrm{tot}(\phi(x)) \longrightarrow \big(\forall x\ [T(\phi(x)) \rightarrow T(\phi(x+1))] \longrightarrow \big(T(\phi(0)) \rightarrow \forall x\ T(\phi(x))\big)\big)\Big),$$

where the quantifier '$\forall \phi(x)$' reads 'for every formula $\phi(x)$ with at most x free'. We read the expression '$\mathrm{tot}(x)$' *as 'x is* total'. *We denote the theory* PT$^-$ + Ind_{tot} *by* $\mathrm{PT_{tot}}$.

[5] Some authors, notably Fischer (2009), have been using the notation "PT$^-$" instead of our PT$_{tot}$. We prefer a different notation, reserving "Th^-" for truth theories Th without any induction for formulae containing the truth predicate.

[6] For better readability, in what follows we will be writing $T(\phi(c))$, instead of $T(\phi(\underline{c}))$. This convention extends also to sentences with quantifiers, e.g. $\forall x T(\phi(x))$ abbreviates $\forall x T(\phi(\underline{x}))$. Note that this applies to the situations in which the occurrence of the predicate T is implicit, as in the formula $\mathrm{tot}(\phi)$.

One could wonder whether such a form of induction is not overly restrictive. Why should we have induction for total formulae only? However, Fischer and Horsten claim that the restriction becomes natural and well motivated as soon as we appreciate that it is *all* models of arithmetic—including the nonstandard ones—that matter:

> We do not want to accept instantiations of induction for nonstandard elements that are not truth-determinate for the property in question for exactly the same reason that we resist inductive premises for soritical predicates. Note that the reply that in the "intended" model there are no such nonstandard elements to be found is not undermining our motivation for restriction; as we have emphasised repeatedly, we are adopting the model-theoretic viewpoint, and from this viewpoint, all models are on a par. (Fischer & Horsten, 2015, p. 355)

In Fischer (2009) it has been claimed that $\mathrm{PT}_{\mathrm{tot}}$ is semantically conservative over PA; in the same paper the author proves that $\mathrm{PT}_{\mathrm{tot}}$ is not interpretable in its base theory (PA). In turn, in Fischer (2014) it is shown that $\mathrm{PT}_{\mathrm{tot}}$ has a nonelementary speed-up over PA. Taken together, these results imply that $\mathrm{PT}_{\mathrm{tot}}$ is indeed an excellent candidate for the role of the theory characterising the use of the truth predicate in model-theoretic contexts.

Unfortunately, the conservativity proof presented in Fischer (2009) contains a flaw.[7] Indeed, the main result of the present paper is that $\mathrm{PT}_{\mathrm{tot}}$ is not semantically conservative over Peano arithmetic (hence the same is true about $\mathrm{PT}_{\mathrm{tot}}$ with (EXT) and (NORM) added). In view of this, the question still remains whether we have at our disposal a natural axiomatic truth theory satisfying requirements (a)–(c).

§2. Nonconservativity of $\mathrm{PT}_{\mathrm{tot}}$**.** An easy argument based on the existence of fixed points for monotone operators shows that PT^- is model-theoretically conservative over PA (see Halbach, 2011). However, it transpires that adding internal induction for total formulae comes with a price. The following theorem states the semantic nonconservativity of $\mathrm{PT}_{\mathrm{tot}}$.

THEOREM 2.1. *There is a model of* PA *which cannot be expanded to a model of* $\mathrm{PT}_{\mathrm{tot}}$.

Before we proceed to the proof, we shall define one construction in propositional logic which is very useful in the context of investigating compositional theories of truth with restricted induction. The construction has originally appeared in Smith (1989) although its properties were not spelled out in full generality.

DEFINITION 2.2. *Let* $\alpha = (\alpha_0 \ldots \alpha_c)$ *and* $\beta = (\beta_0 \ldots \beta_c)$ *be arbitrary sequences of formulae. By a* disjunction of β_i (for $0 \leq k \leq n \leq c$) with stopping condition α*, denoted*

$$\bigvee_{i=k}^{n,\alpha} \beta_i,$$

we mean any of the following formulae defined by backward induction on k:

1. $\bigvee_{i=n}^{n,\alpha} \beta_i = (\alpha_n \wedge \beta_n)$.
2. $\bigvee_{i=k}^{n,\alpha} \beta_i = \neg(\alpha_k \wedge \neg\beta_k) \wedge \Big((\alpha_k \wedge \beta_k) \vee \bigvee_{i=k+1}^{n,\alpha} \beta_i\Big)$.

The intuition behind the above formulae is as follows: we want to search through i as long as we do not see i_0 such that α_{i_0} is true. Then we stop and check whether β_{i_0}

[7] We are grateful to Martin Fischer for the email correspondence concerning this matter.

is satisfied. If it is, then the whole formula is true, if not, then the whole formula is false, regardless of the truth value of β_j for $j > i_0$. We call it a disjunction, since if α is chosen so that exactly one of the α_i-s is true, then the above construction is equivalent in propositional logic to the disjunction:

$$\bigvee_{i=k}^{n} (\alpha_i \wedge \beta_i).$$

Although one should note that in this case it is also equivalent to

$$\bigwedge_{i=k}^{n} (\alpha_i \rightarrow \beta_i).$$

So, in a sense, it is a propositional analogue of a Δ_1-formula.

Given a model M of Peano arithmetic and sequences $\alpha, \beta \in M$,[8] the construction can be reproduced inside M. In such a case, the expression '$\bigvee\limits_{i=k}^{n,\alpha} \beta_i$' will refer to the unique, possibly nonstandard formula in the sense of M. The most important property of the above construction, one we have already alluded to, is that formulae obtained in this way (even nonstandard ones!) behave well in models of compositional truth theories without induction for the language with the truth predicate. These nice properties are encapsulated in the following lemma:

LEMMA 2.3. *Let M be an arbitrary nonstandard model of* PT$^-$. *Let $\alpha, \beta \in M$ be sequences of nonstandard length c, containing as elements arithmetical formulae $\alpha_i(x)$, $\beta_i(y)$ with the free variables indicated.*[9] *Suppose that for all $i \in \omega$ the formulae $\alpha_i(x)$ are standard. Let $a \in M$ be such that $j_0 \in \omega$ is the smallest natural number satisfying the condition $M \models T\big(\alpha_{j_0}(a)\big)$. Then for every $k \leq j_0$ and for every $b \in M$ the following holds*:

(a) $M \models T\Big(\bigvee\limits_{i=k}^{c,\alpha} \beta_i(a,b)\Big) \equiv T\Big(\beta_{j_0}(b)\Big)$.[10]

(b) $M \models T\Big(\neg \bigvee\limits_{i=k}^{c,\alpha} \beta_i(a,b)\Big) \equiv T\Big(\neg\beta_{j_0}(b)\Big)$.

In the proof we will use the well-known fact that for every standard formula $\varphi(x_1 \dots x_n)$:

$$\mathrm{PT}^- \vdash \forall x_1 \dots x_n \Big(T\big(\varphi(x_1 \dots x_n)\big) \equiv \varphi(x_1 \dots x_n)\Big).$$

It immediately follows that for every standard formula $\varphi(x_1 \dots x_n)$:

$$\mathrm{PT}^- \vdash \forall x_1 \dots x_n \Big(T\big(\neg\varphi(x_1 \dots x_n)\big) \equiv \neg T\big(\varphi(x_1 \dots x_n)\big)\Big).$$[11]

[8] More exactly, we assume here that α and β are elements of M such that for some $c \in M$, the formal analogue of the statement 'α and β are sequences of length c containing arithmetical formulae' is true in M.

[9] Again, this is taken to mean that all elements of α and β are formulae in the sense of M.

[10] Note that $\bigvee\limits_{i=k}^{c,\alpha} \beta_i(x,y)$ is a formula with two free variables: the free variable x occurring in the subformulae $\alpha_i(x)$ and the free variable y occurring in the subformulae $\beta_j(y)$.

[11] In effect, we easily obtain the information that for every standard formula $\varphi(x)$, truth is provably total and consistent in PT$^-$. In other words, if $\varphi(x)$ is standard, then PT$^- \vdash \mathrm{tot}(\varphi(x)) \wedge \neg\exists x\Big(T\big(\varphi(x)\big) \wedge T\big(\neg\varphi(x)\big)\Big)$.

Proof. We prove the lemma by backward metainduction on k. Starting with $k = j_0$, note that there exists γ such that the disjunction of β_i with stopping condition α from j_0 to c may be written as follows:

$$\bigvee_{i=j_0}^{c,\alpha} \beta_i(a,b) = \underbrace{\neg(\alpha_{j_0}(a) \wedge \neg\beta_{j_0}(b))}_{A} \wedge \Big(\underbrace{(\alpha_{j_0}(a) \wedge \beta_{j_0}(b)) \vee \gamma}_{B}\Big).$$

For the implication from right to left in part (a) of the lemma, assume that $M \models T(\beta_{j_0}(b))$. Therefore by the axiom of PT$^-$ for double negation, we obtain $M \models T(\neg\neg\beta_{j_0}(b))$, which in turn (by compositional axiom for negated conjunction) permits us to conclude that $M \models T(A)$. Since by assumption $M \models T(\alpha_{j_0}(a))$, we obtain also $M \models T\big(\alpha_{j_0}(a) \wedge \beta_{j_0}(b)\big)$. Applying again compositional axioms of PT$^-$, we get $M \models T(A \wedge B)$, as required.

For the implication from right to left in (b) it is enough to observe that $M \models T(\neg\beta_{j_0}(b))$ together with $M \models T(\alpha_{j_0}(a))$ gives us immediately $M \models T(\neg A)$, and so $M \models T(\neg(A \wedge B))$.

Proving the opposite implication in (a), assume that $M \models T(A \wedge B)$. From $M \models T(A)$ by compositional axioms of PT$^-$ we obtain $M \models T(\neg\alpha_{j_0}(a))$ or $M \models T(\neg\neg\beta_{j_0}(b))$. But $M \models T(\alpha_{j_0}(a))$, and since α_{j_0} is standard, this gives us $M \models \neg T(\neg\alpha_{j_0}(a))$, and so by the compositional axiom for double negation, $M \models T(\beta_{j_0}(b))$.

Proving the opposite implication in (b), assume that $M \models T(\neg(A \wedge B))$, so $M \models T(\neg A) \vee T(\neg B)$. If $M \models T(\neg A)$, then by compositional axioms $M \models T(\neg\beta_{j_0}(b))$. If $M \models T(\neg B)$, then by compositional axioms $M \models T(\neg\alpha_{j_0}(a)) \vee T(\neg\beta_{j_0}(b))$. But $M \models T(\alpha_{j_0}(a))$, so (since α_{j_0} is standard) $M \models \neg T(\neg\alpha_{j_0}(a))$ and therefore $M \models T(\neg\beta_{j_0}(b))$.

Suppose now that our claim is true for a given $k + 1 \leq j_0$. In other words, we have the following:

(i) $M \models T\Big(\bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big) \equiv T\Big(\beta_{j_0}(b)\Big)$.

(ii) $M \models T\Big(\neg \bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big) \equiv T\Big(\neg\beta_{j_0}(b)\Big)$.

Our task is to prove the claim for k. Observe that by Definition 2.2:

$$\bigvee_{i=k}^{c,\alpha} \beta_i = \underbrace{\neg(\alpha_k(a) \wedge \neg\beta_k(b))}_{C} \wedge \underbrace{\Big((\alpha_k(a) \wedge \beta_k(b)) \vee \bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big)}_{D}.$$

For the implication from right to left in part (a) of the lemma, assume that $M \models T(\beta_{j_0}(b))$. Therefore by (i) we have $M \models T\Big(\bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big)$, so $M \models T(D)$. Since $k < j_0$, by the choice of j_0 we have $M \models \neg T(\alpha_k(a))$; therefore $M \models T(\neg\alpha_k(a))$ because α_k is standard. In effect, the compositional axiom of PT$^-$ for negated conjunction permits us to conclude also that $M \models T(C)$ and thus $M \models T(C \wedge D)$.

For the implication from right to left in (b), assume that $M \models T(\neg\beta_{j_0}(b))$. Then by (ii), $M \models T\Big(\neg \bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big)$. Since $M \models \neg T(\alpha_k(a))$, the compositional axioms of PT$^-$ give us $M \models T\big(\neg(\alpha_k(a) \wedge \beta_k(b))\big)$, so $M \models T(\neg D)$ and finally $M \models T\big(\neg(C \wedge D)\big)$ as required.

For the implication from left to right in (a), assume that $M \models T(C \wedge D)$, so in particular $M \models T\big(\alpha_k(a) \wedge \beta_k(b)\big) \vee T\Big(\bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big)$. But $M \models \neg T\big(\alpha_k(a)\big)$ and therefore the second disjunct must be true, which by (i) permits us to conclude that $M \models T\big(\beta_{j_0}(b)\big)$.

For the implication from left to right in (b), assume that $M \models T\big(\neg(C \wedge D)\big)$, so $M \models T(\neg C) \vee T(\neg D)$. However, by compositional axioms $M \models T(\neg C) \equiv \big(T(\alpha_k(a)) \wedge T(\neg\beta_k(b))\big)$ and since $M \models \neg T(\alpha_k(a))$, we obtain $M \models T(\neg D)$. Then by the compositional axiom of PT$^-$ for negated disjunction it follows that $M \models T\Big(\neg \bigvee_{i=k+1}^{c,\alpha} \beta_i(a,b)\Big)$ and therefore by (ii) $M \models T\Big(\neg\beta_{j_0}(b)\Big)$. □

COROLLARY 2.4. *Let M, α, β, a, and j_0 satisfy the assumptions of Lemma 2.3. If in addition the formula $\beta_{j_0}(y)$ is standard, then* $M \models \forall k \leq j_0 \operatorname{tot}\Big(\bigvee_{i=k}^{c,\alpha} \beta_i(a,y)\Big)$.

Proof. It is enough to observe that if $\beta_{j_0}(y)$ is standard, then $M \models \forall b\ T\Big(\big(\beta_{j_0}(b)\big) \vee T\big(\neg\beta_{j_0}(b)\big)\Big)$. Then by Lemma 2.3 the corollary follows immediately. □

Now we are ready to prove our theorem. The argument will consist in showing that no nonstandard prime model of Peano arithmetic is expandable to a model of PT$_{\text{tot}}$.[12]

Proof of Theorem 2.1. Let K be an arbitrary nonstandard prime model of Peano arithmetic. From now on we take for granted that all elements of K are definable in K by arithmetical formulae without parameters. The claim will be that K is not expandable to a model of PT$_{\text{tot}}$. Suppose for contradiction that $(K, T) \models$ PT$_{\text{tot}}$. We will argue that in such a case there is an element e of K which codes $Th(K)$ (that is, e codes the set of all arithmetical sentences true in K). However, such an element e cannot exist in a prime model, because then e would be definable in K, generating a contradiction with Tarski's undefinability theorem.

For starters, fix any nonstandard $c \in K$ and let α be a recursive enumeration of formal definitions up to c. That is, given a fixed recursive enumeration $\phi_0(x), \phi_1(x) \ldots$ of arithmetical formulae with exactly one free variable x, each $\alpha_i(x)$ has the form:

$$\phi_i(x) \wedge \forall y < x \;\; \neg\phi_i(y).$$

For $a \in K$, we say that $\alpha_i(x)$ defines a in K iff i is the smallest natural number such that $K \models \alpha_i(a)$. Observe that since K is prime, such a number exists for every $a \in K$ and a corresponding formula $\alpha_i(x)$ is standard.

We define now the second sequence β, containing formulae $\beta_i(y)$ with one free variable, characterised as follows:

$$\forall \phi < i \left[Sent_{L_{PA}}(\phi) \rightarrow \big(\phi \in y \equiv T_i(\phi)\big)\right].$$

The expression 'T_i' stands for an arithmetical truth predicate for sentences below i. The exact shape of this predicate is not of crucial importance; what will really matter

[12] Given a model M of Peano arithmetic, the universe of a prime model K can be defined as the set of all those elements of M which are definable in M by arithmetical formulae without parameters. In turn, the operations of K are defined as those of M restricted to the universe of K. For further information about prime models we refer the reader to Kaye (1991), p. 91ff.

is that for $i \in \omega$, $T_i(x)$—and therefore also $\beta_i(y)$—is a standard arithmetical formula.[13]

We define

- $\psi(x, y) := \bigvee_{i=0}^{c,\alpha} \beta_i(x, y)$,
- $\xi(z) := \exists y \forall x < z\ \psi(x, y)$.

We emphasise that $\psi(x, y)$ and $\xi(z)$ are defined in a model K and not externally. With c being nonstandard, both $\psi(x, y)$ and $\xi(z)$ should be thought of as nonstandard arithmetical formulae—elements of K which are *perceived* by K as formulae, not to be confused with the 'real world' arithmetical expressions.[14]

We are going to show that

(i) $(K, T) \models \mathrm{tot}\big(\xi(z)\big)$,

(ii) $(K, T) \models T\big(\xi(0)\big) \wedge \forall x \Big(T\big(\xi(x)\big) \rightarrow T\big(\xi(x + 1)\big)\Big)$.

For (i), first observe that $(K, T) \models \forall a\ \mathrm{tot}(\psi(a, y))$. In order to see this, fix $a \in K$ and let $\alpha_{j_0}(x)$ define a. Fixing b, it is enough to observe that the assumptions of Corollary 2.4 are satisfied; therefore $(K, T) \models \forall k \leq j_0\ \mathrm{tot}\Big(\bigvee_{i=k}^{c,\alpha} \beta_i(a, y)\Big)$. For $k = 0$, this gives $(K, T) \models \forall a\ \mathrm{tot}(\psi(a, y))$.

It easily follows that $\forall a \in K\ (K, T) \models \mathrm{tot}\big(\forall x < a \psi(x, y)\big)$.[15] Finally, we argue for the totality of $\xi(z)$. Given an arbitrary $a \in K$, if $(K, T) \models \exists y T(\forall x < a\ \psi(x, y))$, then $(K, T) \models T(\xi(a))$. Otherwise $(K, T) \models \forall y \neg T(\forall x < a\ \psi(x, y))$, which (by totality of '$\forall x < a\ \psi(x, y)$') entails $(K, T) \models T(\neg\xi(a))$.

For (ii), obviously $(K, T) \models T\big(\xi(0)\big)$,[16] so we move to the second conjunct. Fix b such that $(K, T) \models T\big(\xi(b)\big)$; in other words, $(K, T) \models T\big(\exists y \forall x < b\ \psi(x, y)\big)$. Choose $e \in K$ such that $(K, T) \models T\big(\forall x < b\ \psi(x, e)\big)$. Our task is to obtain $e' \in K$ such that $(K, T) \models T\big(\forall x < b + 1\ \psi(x, e')\big)$.

Let $\alpha_i(x)$ be the least definition of b. We put

$$e' = e - \{0 \ldots i - 1\} \cup \{\psi \in Sent_{L_{PA}} \mid \psi < i \wedge K \models \psi\}.$$

All that remains for the completion of the proof of (ii) is to show that for every $d < b+1$, $(K, T) \models T\big(\psi(d, e')\big)$. Fix such a d and let $\alpha_j(x)$ be the least definition of d. By Lemma 2.3, it is enough to obtain $(K, T) \models T\big(\beta_j(e')\big)$; in other words, we want to have

[13] For example, predicates $T_i(x)$ could be defined as '$(x = \ulcorner\psi_0\urcorner \wedge \psi_0) \vee \cdots \vee (x = \ulcorner\psi_m\urcorner \wedge \psi_m)$', where $\psi_0 \ldots \psi_m$ are all arithmetical sentences with Gödel numbers smaller than i. Note that this construction of $T_i(x)$ can be carried out in an arbitrary model M of PA. In effect, we will have formulae (in the sense of M) $T_c(x)$ for an arbitrary element c of M, even a nonstandard one. However, it is important to emphasise that in each formula $T_i(x)$ the letter 'i' is not a variable. The index gives rather the information that in the corresponding formula all sentences with Gödel numbers smaller than (fixed) i are taken into account.

[14] Accordingly, it would make no sense to say, for example, that $\psi(x, y)$ is satisfied in K by some objects a and b. Nevertheless, nonstandard formulae can be employed in the scope of the truth predicate and we are allowed to say, for example, that $(K, T) \models T(\psi(a, b))$.

[15] Fix a and b; assume that $(K, T) \models \neg T\big(\forall x < a\ \psi(x, b)\big)$. Then $(K, T) \models \exists x < a \neg T\big(\psi(x, b)\big)$. Choosing such an $x \in K$, we obtain $(K, T) \models \neg T\big(\psi(x, b)\big)$ and by totality of $\psi(a, y)$ for every a, we obtain $(K, T) \models T\big(\neg\psi(x, b)\big)$ and thus $(K, T) \models T\big(\neg\forall x < a\ \psi(x, b)\big)$.

[16] It is enough to observe that there is no $x < 0$ in K, so an arbitrary y is a witness for '$\exists y \forall x < 0\ \psi(x, y)$'.

$$(*)\ \ (K,T) \models T\Big(\forall\phi < j\big[Sent_{L_{PA}}(\phi) \rightarrow \big(\phi \in e' \equiv T_j(\phi)\big)\big]\Big).$$

Since $d < b+1$, we consider two cases. If $d = b$, then $j = i$ and (*) follows trivially from the definition of e'. On the other hand, if $d < b$, then $(K,T) \models T\big(\psi(d,e)\big)$, and so by Lemma 2.3 $(K,T) \models T\big(\beta_j(e)\big)$. In other words, we have

$$(**)\ \ (K,T) \models T\Big(\forall\phi < j\big[Sent_{L_{PA}}(\phi) \rightarrow \big(\phi \in e \equiv T_j(\phi)\big)\big]\Big).$$

It is easy now to show that (*) must hold. Fixing $\phi < j$, we observe that if $\phi < i$, then $(K,T) \models T\big(\phi \in e' \equiv T_j(\phi)\big)$ because by definition, e' codes only true sentences below i. Otherwise $\phi \geq i$, but then $\phi \in e$ and $\phi \in e'$, so by (**) we also conclude that $(K,T) \models T\big(\phi \in e' \equiv T_j(\phi)\big)$.

This finishes the proofs of (i) and (ii). At this point we know that $\xi(z)$ is total and inductive in (K,T), so by the axiom of internal induction we conclude that $(K,T) \models \forall z\ T(\xi(z))$. Let a be a nonstandard element of K. Since $(K,T) \models T(\xi(a))$, we can choose e such that $(K,T) \models \forall x < a\ T\big(\psi(x,e)\big)$.

We will show that e codes $Th(K)$ in K; in other words, we show that

$$(***)\ \ \forall n \in \omega\ K \models \forall\phi < n\ \big[Sent_{L_{PA}}(\phi) \rightarrow \big(\phi \in e \equiv T_n(\phi)\big)\big].$$

Fix $n \in \omega$. Let k and i be elements of ω such that $\alpha_i(x)$ defines k in K and $i > n$. Since $k < a$,[17] we have $(K,T) \models T\big(\psi(k,e)\big)$, which by Lemma 2.3 is equivalent to $(K,T) \models T\big(\beta_i(e)\big)$. In other words,

$$(K,T) \models T\Big(\forall\phi < i\ \big[Sent_{L_{PA}}(\phi) \rightarrow \big(\phi \in e \equiv T_i(\phi)\big)\big]\Big).$$

Applying disquotation (valid in PT$^-$ for standard formulae with parameters), we obtain

$$K \models \forall\phi < i\ \big[Sent_{L_{PA}}(\phi) \rightarrow \big(\phi \in e \equiv T_i(\phi)\big)\big].$$

Since $i > n$, (***) follows trivially. In effect, e codes $Th(K)$ in K. But this is impossible in prime models, thus a contradiction is obtained and the proof is finished. □

§3. Expandability properties of models of PA.

In this section we further approximate the class of those models of PA that admit an expansion to models of PT$_{\text{tot}}$. In general, such a class is a handy tool for comparing properties of axiomatic theories of truth. Let us introduce the precise definition and a piece of notation:

DEFINITION 3.1. *Let Th be any extension of* PA *(possibly in extended language). By $\mathfrak{Th}$ we denote the class of models of* PA *that can be expanded to a model of Th.*

For example, it is known that for every nonstandard model $\mathcal{M}$, $\mathcal{M} \in \mathfrak{TB}$ if and only if $\mathcal{M}$ codes its own theory, i.e.,

$$\{\ulcorner\phi\urcorner \mid \phi \in Sent_{L_{PA}} \wedge \mathcal{M} \models \phi\}$$

is coded in $\mathcal{M}$.[18]

[17] We remind that k is standard and a is a nonstandard element of K.

[18] For the details, see Cieśliński (2015b). However, it should be noted that our ability to provide a purely arithmetical characterisation of $\mathfrak{TB}$ makes TB rather exceptional. For most axiomatic truth theories we are only able to restrict the class of possible candidates. For example, it is known that the class of recursively saturated models of PA properly contains $\mathfrak{CT}^-$ and $\mathfrak{UTB}$.

Results on such classes can be used to obtain some information about *relative truth definability*[19] between axiomatic theories of truth. The connection between these two notions is established via the following fact:

FACT 3.2. *Let Th_1, Th_2 be two axiomatic truth theories. If Th_1 is relatively truth definable in Th_2, then $\mathfrak{Th}_2 \subseteq \mathfrak{Th}_1$.*

For the (immediate) proof, see Fujimoto (2010).

In the previous section we showed that the class of prime models of PA is disjoint from $\mathfrak{PT}_{\mathfrak{tot}}$. In the next theorem, we approximate this class from below (by $\mathfrak{RS}$ we denote the class of recursively saturated models of PA). For technical simplicity, we will prove the theorem for the stronger theory $\mathrm{PT}_{\mathrm{tot}}$ + (EXT) + (NORM), which from now on we will denote by $\mathrm{RPT}_{\mathrm{tot}}$ (*regular* $\mathrm{PT}_{\mathrm{tot}}$).

THEOREM 3.3. $\mathfrak{RS} \subseteq \mathfrak{RPT}_{\mathfrak{tot}}$.[20]

Note that since $\mathrm{PT}_{\mathrm{tot}}$ is a subtheory of $\mathrm{RPT}_{\mathrm{tot}}$, the above result clearly implies that every recursively saturated model of PA can be expanded to a model of $\mathrm{PT}_{\mathrm{tot}}$.

The proof follows immediately from two lemmata, which we consider interesting also for their own sake. Before stating them, let us introduce two more notions (the expressions $Sent_{\mathcal{M}}$, $Tm^c_{\mathcal{M}}$, $Form^1_{\mathcal{M}}$ denote the set of arithmetical sentences, closed terms and arithmetical formulae with at most one free variable in the sense of a model $\mathcal{M}$):

DEFINITION 3.4. *By the* term formulation *of* PT^-, *denoted by* tPT^-, *we mean* PT^- *with the quantifier axioms*:

$$\text{(6)}\ \forall v \in Var \forall \varphi \in L_{PA}\big(T(\forall v\varphi) \equiv \forall x T(\varphi(\underline{x}/v))\big)$$

$$\text{(7)}\ \forall v \in Var \forall \varphi \in L_{PA}\big(T(\neg\forall v\varphi) \equiv \exists x T(\neg\varphi(\underline{x}/v))\big)$$

replaced with the following ones:

$$\text{(6')}\ \forall v \in Var \forall \varphi \in L_{PA}\big(T(\forall v\varphi) \equiv \forall t \in Tm^c T(\varphi(t/v))\big)$$

$$\text{(7')}\ \forall v \in Var \forall \varphi \in L_{PA}\big(T(\neg\forall v\varphi) \equiv \exists t \in Tm^c T(\neg\varphi(t/v))\big).$$

We define the term formulation of $\mathrm{PT}_{\mathrm{tot}}$ *as the term formulation of* PT^- *extended with the internal induction axiom* (Ind_{tot}). *We note that there is no need to define the term formulations of* RPT *or* $\mathrm{RPT}_{\mathrm{tot}}$, *since over any extension of* PT^- *containing* (EXT) *the conditions* 6 (7) *and* 6′ (7′) *are equivalent.*

DEFINITION 3.5. *Let $\mathcal{M} \models$ PA, let $A \subseteq M$ and let ϕ be a L_{PA} sentence in the sense of $\mathcal{M}$. We define*

$$\begin{aligned}\Theta_{\mathcal{M}}(\phi, A) := {} & \mathcal{M} \models \exists s, t \in Tm^c[\phi = (s = t) \wedge val(s) = val(t)]\\ & \vee\ \mathcal{M} \models \exists s, t \in Tm^c[\phi = \neg(s = t) \wedge val(s) \neq val(t)]\\ & \vee\ \exists \psi \in Sent_{\mathcal{M}}[\mathcal{M} \models \phi = \neg\neg\psi] \wedge \psi \in A\\ & \vee\ \exists \psi_1, \psi_2 \in Sent_{\mathcal{M}}[\mathcal{M} \models \phi = (\psi_1 \wedge \psi_2)] \wedge \big(\psi_1 \in A \wedge \psi_2 \in A\big)\end{aligned}$$

[19] For the definition of this notion and the discussion of its philosophical relevance, see Fujimoto (2010).

[20] Cf. Cantini (1989), where it is argued that every recursively saturated model is expandable to a model of an untyped truth theory KF_{t}. Cantini's KF_{t} is formulated in the language with two predicates 'T' and 'F' (for truth and falsity respectively); it contains also the axiom of internal induction for total formulae.

$$
\begin{aligned}
&\vee\ \exists\psi_1,\psi_2 \in Sent_{\mathcal{M}}[\mathcal{M}\models\phi = \neg(\psi_1\wedge\psi_2)] \wedge \big((\neg\psi_1\in A)\vee(\neg\psi_2\in A)\big)\\
&\vee\ \exists\psi_1,\psi_2 \in Sent_{\mathcal{M}}[\mathcal{M}\models\phi = (\psi_1\vee\psi_2)] \wedge \big((\psi_1\in A)\vee(\psi_2\in A)\big)\\
&\vee\ \exists\psi_1,\psi_2 \in Sent_{\mathcal{M}}[\mathcal{M}\models\phi = \neg(\psi_1\vee\psi_2)] \wedge \big((\neg\psi_1\in A)\wedge(\neg\psi_2\in A)\big)\\
&\vee\ \exists\psi(x) \in Form^1_{\mathcal{M}}[\mathcal{M}\models\phi = \exists x\psi] \wedge \exists s\in Tm^c\ \ (\psi(s)\in A)\\
&\vee\ \exists\psi(x) \in Form^1_{\mathcal{M}}[\mathcal{M}\models\phi = \neg\exists x\psi] \wedge \forall s\in Tm^c\ \ (\neg\psi(s)\in A)\\
&\vee\ \exists\psi(x) \in Form^1_{\mathcal{M}}[\mathcal{M}\models\phi = \forall x\psi] \wedge \forall s\in Tm^c\ \ (\psi(s)\in A)\\
&\vee\ \exists\psi(x) \in Form^1_{\mathcal{M}}[\mathcal{M}\models\phi = \neg\forall x\psi] \wedge \exists s\in Tm^c\ \ (\neg\psi(s)\in A).
\end{aligned}
$$

Let $\Gamma^M : \mathcal{P}(M)\to\mathcal{P}(M)$ *be the function defined in the following way*:

$$\Gamma^M(A) = \{\phi\in\mathcal{M}\ \mid\ \Theta_{\mathcal{M}}(\phi, A)\}. \qquad (\Gamma)$$

As usual, any set $P\subseteq M$ satisfying $\Gamma^M(P) = P$ will be called *a fixpoint of* Γ^M. Such functions generate the extension for the truth predicate of tPT$^-$ in an arbitrary model of PA. More precisely, we have the following well-known fact:

FACT 3.6. *Let* $\mathcal{M}\models$ PA. *Then*

(1) *the operator* Γ^M *is monotone with respect to inclusion, that is, for every sets* $A\subseteq B\subseteq M$ *we have*

$$\Gamma^M(A)\subseteq\Gamma^M(B),$$

(2) *there exists a fixpoint of* Γ^M,

(3) *if* $P\subseteq M$ *is any fixpoint of* Γ^M, *then* $(\mathcal{M}, P)\models tPT^-$.

Our proof of Theorem 3.3 will be based on the existence of particularly well-behaved fixpoints of Γ in recursively saturated models.

DEFINITION 3.7. *Let* $\mathcal{M}\models$ PA.

$$
\begin{aligned}
\Gamma^M_0 &= \emptyset,\\
\Gamma^M_{n+1} &= \Gamma^M(\Gamma^M_n),\\
\Gamma^M_\omega &= \bigcup_{n\in\omega}\Gamma^M_n.
\end{aligned}
$$

Convention. Instead of Γ^M $(\Gamma^M_n, \Gamma^M_\omega)$, we will write Γ $(\Gamma_n, \Gamma_\omega)$ omitting the upper index indicating the model.

The following proposition witnesses the first two nice properties of Γ_ω—*if it is a fixpoint of* Γ, *then it is consistent and satisfies both the extensionality and the normality principles*.[21]

PROPOSITION 3.8. *Let* $\mathcal{M}\models$ PA *and suppose that* Γ_ω *is a fixpoint of* Γ *in* M. *Then*

$$(\mathcal{M},\Gamma_\omega)\models \text{RPT}^- + \forall\phi\in Sent_{L_{PA}}\big(\neg(T\phi\wedge T\neg\phi)\big).$$

Proof. Fix $\mathcal{M}\models$ PA and suppose Γ_ω is a fixpoint of Γ. Then by Fact 3.6,

$$(\mathcal{M},\Gamma_\omega)\models t\text{PT}^-.$$

[21] In fact this proposition is more general: one can show (by a trivial generalisation of our proof) that any least fixpoint of Γ is consistent. Since we do not need it in full generality, we prove the particular case only.

That (NORM) holds in $(\mathcal{M}, \Gamma_\omega)$ is straighforward from the definition of Γ. We show consistency first, i.e., we show that for no $\phi \in Sent_{\mathcal{M}}$ it holds that

$$(\mathcal{M}, \Gamma_\omega) \models T(\phi \wedge \neg\phi).$$

Let us observe that ϕ cannot be an atomic sentence, since PT$^-$ proves that truth is consistent for atomic sentences. Suppose the above is false and take the least $n \in \omega$ such that

$$(\mathcal{M}, \Gamma_n) \models T(\phi \wedge \neg\phi)$$

for some *nonatomic* $\phi \in Sent_{\mathcal{M}}$. Then $n \neq 0$ since Γ_0 is empty. By considering all possible grammatical forms of ϕ we show that for some $\psi \in Sent_{\mathcal{M}}$

$$(\mathcal{M}, \Gamma_{n-1}) \models T(\psi \wedge \neg\psi),$$

which contradicts the choice of n.[22]

Let us now handle (EXT). Suppose for some $t, s \in Tm^c_{\mathcal{M}}$, $\phi(x) \in Sent_{\mathcal{M}}$ such that $\mathcal{M} \models val(s) = val(t)$ we have

$$(\mathcal{M}, \Gamma_\omega) \models T(\phi(s)) \wedge \neg T(\phi(t)).$$

Let n be the least number such that for some formula $\psi(x)$, $\psi(s) \in \Gamma_n$ and $\psi(t) \notin \Gamma_n$. Once again observe that ψ cannot be atomic, since axiom (1) of PT$^-$ guarantees extensionality for such sentences. Moreover $n \neq 0$, since we assumed Γ_n to be nonempty. By considering all the grammatical forms of ψ we show that there is a formula $\theta(x)$ such that

$$(\mathcal{M}, \Gamma_{n-1}) \models T(\theta(s)) \wedge \neg T(\theta(t)).$$

Let us do the step for $\vee$. If $\psi(x) = \theta_1(x) \vee \theta_2(x)$, then by the assumption concerning $\psi(t)$, for every $k \leq n$,

$$\theta_1(t) \notin \Gamma_k \text{ and } \theta_2(t) \notin \Gamma_k.$$

Since $\psi(s) \in \Gamma_n$, we have that either $\theta_1(s) \in \Gamma_{n-1}$ or $\theta_2(s) \in \Gamma_{n-1}$, which contradicts the choice of n. □

Let us state our two promised lemmata:

LEMMA 3.9. *If $\mathcal{M} \models$ PA is recursively saturated, then Γ_ω is a fixpoint of Γ.*

LEMMA 3.10. *For every $\mathcal{M} \models$ PA, if Γ_ω is a fixpoint of Γ, then $(\mathcal{M}, \Gamma_\omega) \models \mathrm{RPT_{tot}}$.*

Proof of Lemma 3.9. Let $\mathcal{M} \models$ PA be recursively saturated. We have to check that $\Gamma(\Gamma_\omega) = \Gamma_\omega$. Since Γ is monotone it is sufficient to show that

$$\Gamma(\Gamma_\omega) \subseteq \Gamma_\omega.$$

Let $\phi \in \Gamma(\Gamma_\omega)$. This means that $\Theta_{\mathcal{M}}(\phi, \Gamma_\omega)$. By considering all the disjuncts constituting $\Theta_{\mathcal{M}}(\phi, \Gamma_\omega)$, we show that $\phi \in \Gamma_\omega$. The only nontrivial steps where we use the recursive saturation of $\mathcal{M}$ are hidden in the cases when $\phi = \forall x \psi$ and $\phi = \neg\exists x \psi$. Let us do the former, the proof of the latter being fully analogous. We have to check that

$$\forall x \psi \in \Gamma_\omega \text{ if and only if for every } t \in Tm^c_{\mathcal{M}} \;\; \psi(t) \in \Gamma_\omega. \tag{1}$$

From left to right, this is immediate: if $\forall x \psi \in \Gamma_\omega$, then there exists $n \in \omega$ such that $\forall x \psi \in \Gamma_n$. Then, by the definition of Γ it has to be the case that for every $t \in Tm^c_{\mathcal{M}}$,

$$\psi(t) \in \Gamma_{n-1}.$$

[22] For example, if $\phi = \exists x \psi(x)$, then by the definition of Γ_n there must be a number $s \in \mathcal{M}$ such that $(\mathcal{M}, \Gamma_{n-1}) \models T\big(\psi(s) \wedge \neg\psi(s)\big)$.

Since $\Gamma_{n-1} \subseteq \Gamma_\omega$, then the same holds also for Γ_ω (this is also the general pattern of proving the steps for sentential connectives). This ends the proof of the first implication. Suppose now that for every $t \in Tm^c_{\mathcal{M}}$,

$$\psi(t) \in \Gamma_\omega. \tag{$*$}$$

We claim that there is a number $n \in \omega$ such that for every $t \in Tm^c_{\mathcal{M}}$,

$$\psi(t) \in \Gamma_n. \tag{$**$}$$

Aiming at a contradiction, suppose the contrary. Observe that the condition defining Γ_n can be written as an arithmetical formula $\Gamma_n(x)$ such that for every $\phi \in Sent_{\mathcal{M}}$,

$$\phi \in \Gamma_n \text{ if and only if } \mathcal{M} \models \Gamma_n(\phi).^{23}$$

Let us consider the following set:

$$p(x) = \{x \in Tm^c \wedge \neg\Gamma_n(\psi(x)) \mid n \in \omega\}.$$

From our assumption that $(**)$ does not hold, together with the fact that for every $k < l \in \omega$,

$$\mathcal{M} \models \forall x \big(\Gamma_k(x) \rightarrow \Gamma_l(x)\big),$$

we conclude that $p(x)$ is finitely realisable in $\mathcal{M}$, so it is a recursive type. By recursive saturation there exists $t \in M$ realising $p(x)$. It follows that $t \in Tm^c_{\mathcal{M}}$ and $\psi(t) \notin \Gamma_\omega$, so we obtain a contradiction with $(*)$.

Now, let $n \in \omega$ be any number satisfying $(**)$. Then by definition of Γ we have

$$\forall x \psi(x) \in \Gamma_{n+1}$$

and therefore

$$\forall x \psi(x) \in \Gamma_\omega,$$

which ends the proof. □

Proof of Lemma 3.10. Since we know that Γ_ω is a fixpoint of Γ, by Proposition 3.8 we conclude that

$$(\mathcal{M}, \Gamma_\omega) \models \mathrm{RPT}^-$$

and Γ_ω is consistent. In effect, it is sufficient to show that $(\mathcal{M}, \Gamma_\omega)$ satisfies the internal induction for total formulae. Fix a formula $\phi(x)$ and suppose that

$$(\mathcal{M}, \Gamma_\omega) \models \forall t \in Tm^c \big(T(\phi(t)) \vee T(\neg\phi(t))\big).^{24}$$

Then by compositional axioms of PT$^-$ we have $(\mathcal{M}, \Gamma_\omega) \models T\big(\forall x(\phi(x) \vee \neg\phi(x))\big)$. Hence there is a number $n \in \omega$ such that

$$\forall x(\phi(x) \vee \neg\phi(x)) \in \Gamma_n.$$

It follows that for every $t \in Tm^c_{\mathcal{M}}$ $\phi(t) \in \Gamma_n \vee \neg\phi(t) \in \Gamma_n$. Since Γ_ω is consistent, we conclude that for every $t \in Tm^c_{\mathcal{M}}$

$$\psi(t) \in \Gamma_n \text{ if and only if } \psi(t) \in \Gamma_\omega.$$

[23] We define $\Gamma_n(x)$ inductively: $\Gamma_0(x) := x \neq x$ and $\Gamma_{n+1}(x) := \Theta(x, \Gamma_n)$, where '$\Theta(x, \Gamma_n)$' is the arithmetical formula obtained from '$\Theta_{\mathcal{M}}(\phi, \Gamma_n)$' by omitting the mention of the model.

[24] Since $(\mathcal{M}, \Gamma_\omega)$ satisfies RPT$^-$, it makes (EXT) true; therefore not only the compositional axioms for quantifiers, but also the totality condition can be equivalently formulated in its present version, employing the quantification over closed terms.

But Γ_n is *arithmetically definable* by formula $\Gamma_n(x)$, so we have the induction axiom for it already in PA. □

Theorem 3.3 follows easily from Lemmas 3.9 and 3.10.

REMARK 3.11. We have decided to isolate the two lemmata, because we believe that they both give rise to interesting questions. The first one, based on Lemma 3.9, is as follows: *What can be said about a model $\mathcal{M}$ given that we know the number of iterations of Γ needed to obtain the least fixpoint?* What we do know is that the converse to Lemma 3.9 holds, i.e., if Γ_ω is a fixpoint of Γ, then $\mathcal{M}$ is recursively saturated (the proof of this observation is beyond the scope of this paper). The second question, based on Lemma 3.10, is as follows: *What other fixpoints of Γ give rise to interpretations of* $\mathrm{PT}_{\mathrm{tot}}$*?* Here we have no clear intuitions.

Let us note one immediate corollary to Theorem 3.3:

COROLLARY 3.12. CT^- *and* UTB *are not relatively truth definable in* $\mathrm{PT}_{\mathrm{tot}}$.

Proof. By theorems of Kaufmann–Schmerl (see Kossak & Schmerl, 2006) and Smith (1989), there is a recursively saturated model of PA which does not admit an expansion to a model of CT^- or UTB (a *rather classless* model). By Theorem 3.3, this model (being recursively saturated) can be expanded to a model of $\mathrm{PT}_{\mathrm{tot}}$. Hence, there is a model of PA which expands to a model of $\mathrm{PT}_{\mathrm{tot}}$ but does not expand to a model of CT^- or UTB. Our corollary follows now from Fact 3.2. □

§4. Concluding remarks.

By Theorem 2.1, we know that $\mathrm{PT}_{\mathrm{tot}}$ does not satisfy the semantic conservativity condition. On the other hand, by Theorem 3.3 we also know that, unlike in the case of CT^-, the interpretation of the truth predicate of $\mathrm{PT}_{\mathrm{tot}}$ can be found in an arbitrary recursively saturated model of Peano arithmetic. Now, where does it leave Fischer's and Horsten's project?

In view of the aforementioned results, there are two possible moves to consider. One of them consists in rejecting $\mathrm{PT}_{\mathrm{tot}}$ and searching for another truth theory, satisfying all the demands (a)–(c) from p. 188 of this paper. The second option would involve modifying the demands, in particular, weakening the semantic conservativity requirement.

It is our opinion that the adoption of the second strategy would have far-reaching consequences; namely, that it would amount to nothing less than dropping the original project of providing a characterisation of the use of the truth predicate in model theory. As stressed by Fischer and Horsten, general model theory does not discriminate between models and any description of the notion of model-theoretic truth should take this fact into account. Why then should expandability of all recursively saturated models—but not, say, of prime models—matter for such an endeavour? What is it that permits us to treat model theory as being specifically *about* recursively saturated models? We are not aware of any convincing answer to this question.

Admittedly, expandability of recursively saturated models guarantees the syntactic conservativity of a given theory of truth. Indeed, syntactic conservativity is crucial for a related but different philosophical project, proposed in Fischer (2014) and Fischer (2015), where the core idea is to present truth as an instrumental device, on a par with 'ideal elements in mathematics' in Hilbert's programme (cf. Fischer, 2015, p. 294). In such a context, there is still a place for theories like $\mathrm{PT}_{\mathrm{tot}}$. Nevertheless, we should emphasise that the present project—that of characterising the use of the truth predicate in model theory—is different and we just do not see how it could proceed without the semantic conservativity condition.

In effect, from the philosophical point of view the first option definitely seems more attractive. Moreover, some axiomatic truth theories in the vicinity of $\mathrm{PT}_{\mathrm{tot}}$ merit further attention as very promising candidates for the role of a theory realising Fischer's and Horsten's postulates.[25]

Acknowledgment. The research presented in this paper was supported by the National Science Centre, Poland (NCN), grant number 2014/13/B/HS1/02892.

BIBLIOGRAPHY

Cantini, A. (1989). Notes on formal theories of truth. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, **35**(2), 97–130.

Cieśliński, C. (2015a). The innocence of truth. *Dialectica*, **69**(1), 61–85.

Cieśliński, C. (2015b). Typed and untyped disquotational truth. In Achourioti, T., Galinon, H., Fujimoto, K., and Martínez-Fernández, J., editors. *Unifying the Philosophy of Truth*. Dordrecht: Springer, pp. 307–320.

Cieśliński, C. (2011). T-equivalences for positive sentences. *The Review of Symbolic Logic*, **4**(2), 319–325.

Fischer, M. (2009). Minimal truth and interpretability. *The Review of Symbolic Logic*, **2**(04), 799–815.

Fischer, M. (2014). Truth and speed-up. *The Review of Symbolic Logic*, **7**(2), 319–340.

Fischer, M. (2015). Deflationism and instrumentalism. In Achourioti, T., Galinon, H., Fujimoto, K., and Martínez-Fernández, J., editors. *Unifying the Philosophy of Truth*. Dordrecht: Springer, pp. 293–306.

Fischer, M. & Horsten, L. (2015). The expressive power of truth. *The Review of Symbolic Logic*, **8**(2), 345–369.

Fujimoto, K. (2010). Relative truth definability of axiomatic truth theories. *The Bulletin of Symbolic Logic*, **16**, 305–344.

Halbach, V. (2009). Reducing compositional to disquotational truth. *The Review of Symbolic Logic*, **2**(4), 786–798.

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

Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In Cartois, P., editor. *The Many Problems of Realism*. Studies in the General Philosophy of Science, Vol. 3. Tilburg: Tilburg University Press, pp. 173–187.

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

Ketland, J. (1999). Deflationism and Tarski's paradise. *Mind*, **108**(429), 69–94.

Kossak, R. & Schmerl, J. H. (2006). *The Structure of Models of Peano Arithmetic*. Oxford: Clarendon Press.

Shapiro, S. (1998). Proof and truth: Through thick and thin. *The Journal of Philosophy*, **95**(10), 493–521.

Smith, S. T. (1989). Nonstandard definability. *Annals of Pure and Applied Logic*, **42**, 21–43.

[25] One of these promising theories is $\mathrm{WPT}_{\mathrm{tot}}$, which is obtained from $\mathrm{PT}_{\mathrm{tot}}$ by modifying the compositional truth axioms in accordance with the principles of Weak Kleene logic. However, the analysis of the properties of $\mathrm{WPT}_{\mathrm{tot}}$ and related theories goes beyond the scope of the present paper.

INSTITUTE OF PHILOSOPHY
UNIVERSITY OF WARSAW
WARSAW, POLAND
E-mail: c.cieslinski@uw.edu.pl
E-mail: mlelyk@student.uw.edu.pl
E-mail: bar.wcislo@gmail.com