Bartosz Wcisło[1], **Mateusz Łełyk**[2]

Strong and Weak Truth Principles[3]

Abstract This paper is an exposition of some recent results concerning various notions of strength and weakness of the concept of truth, both published or not. We try to systematically present these notions and their relationship to the current research on truth. We discuss the concept of the Tarski boundary between weak and strong theories of truth and we give an overview of non-conservativity results for the extensions of the basic compositional truth theory. Additionally, we present a natural strong theory of truth which admits a number of apparently unrelated axiomatisations. Finally, we discuss other possible explications of the notion of 'strength' of axiomatic theories of truth.

Keywords Axiomatic truth theories, Peano arithmetic, Conservativity, Tarski boundary

1. Introduction

1.1 Axiomatic theories of truth

Formal theories of truth are a part of philosophy investigating the notion of truth with the methods of mathematical logic. One of the main methods of formalisation is to consider *axiomatic theories of truth* which are constructed in the following way:

[1]University of Warsaw, Faculty of Mathematics, Informatics and Mechanics.
E-mail: b.wcislo@mimuw.edu.pl
[2]University of Warsaw, Faculty of Philosophy and History.
E-mail:mlelyk@student.uw.edu.pl
[3]Research reported in this paper was conducted with a financial support of the Polish National Science Centre, Grant number 2014/13/B/HS1/02892.

- We fix a **base theory** B modelling the totality of our knowledge of extra-semantic facts (facts not concerning such notions as meaning or truth).

- We add to the language of that theory a new unary predicate $T(x)$ with the intended reading "x is a true sentence" and we extend B with axioms governing the new predicate.

We then investigate, how the properties of the obtained theory depend on the choice of axioms governing the truth predicate.

The base theory B is often chosen to be Peano arithmetic PA. The motivation behind this choice is that the vast majority of results concerning the relationship between truth theory and its corresponding base theory do not significantly depend on the specific choice of the latter. The only thing which we do require is that it is capable of expressing and proving basic facts concerning syntax, like: "every sentence which is built correctly has the same number of right and left parentheses." PA is more than enough to this end.

It is worth stressing that most logicians do not think of PA as a theory of numbers but rather as a more general theory of finite mathematical objects like hereditarily finite sets, finite graphs, finite strings of characters over a finite alphabet. This theory suffices to prove surprisingly many facts concerning these kinds of objects[4]. Since sentences, formulae or proofs in formal languages may also be treated as finite mathematical objects, namely strings of characters with some simple structural properties, Peano arithmetic allows us to freely speak about them. Having said that, we have to admit that the choice of PA as a base theory is somewhat arbitrary. Formulating in an abstract way the conditions guaranteeing that a base theory is strong enough from the point of view of truth theory, and which suffice to prove the results which make PA our choice, seems a rather daunting task. At this initial stage of research, we prefer the "bottom-up" strategy.

In this paper, we focus on theories describing the truth predicate for the language of the base theory. However, let us stress that the properties of self-referential truth predicates which formalise the notion of truth for all sentences of the language to which they belong are a subject of extensive studies (a good account of results concerning such theories may be found in (Halbach, 2011)).

[4]There are many sources concerning formalisation of syntax and making above comments precise. We especially recommend (Franzen, 2003).

The role of the axioms governing the truth predicate is, obviously, to capture various intuitions concerning this notion. In the studies on formal truth theories, we are trying to explain what are the relations between those intuitive properties, and what are the consequences of the fact that the predicate enjoys these features. One of the simplest conditions for the truth predicate which we may consider is as follows:

$$\varphi \equiv T(\varphi)$$

for all sentences in the language of the base theory. These axioms say that the truth predicate satisfies Tarski's biconditionals for the language of the base theory. The theory extending PA in which the only axioms governing the truth predicate are these biconditionals is called TB^{-}[5].

Another property, which should be satisfied by the truth predicate is compositionality. We express it with axioms formalising principles such as:

> (For all sentences φ and ϑ) The conjunction of sentences φ and ϑ from the language of the base theory is true if and only if both of the conjuncts are.

Or:

> (For an arbitrary variable v and an arbitrary formula $\varphi(v)$ with at most one free variable v) The universal sentence $\forall\varphi(v)$ in the language of the base theory is true if and only if for any numeral $\underline{x}$, the sentence $\varphi(\underline{x})$ is true.
>
> Let us add that by a numeral $\underline{x}$, we mean the canonical term denoting the number x, for instance $(\ldots((\overline{0}+\overline{1})+\overline{1})\ldots+\overline{1})$, where the addition symbol occurs x times (and where $\overline{0}$ and $\overline{1}$ are some fixed symbols representing 0 and 1, respectively).

Let us observe that already the theory TB^{-} can prove for any two concrete sentences from the language of arithmetic that their conjunction[6] is true if and only if both are true. However, it cannot prove the general fact about all arithmetical sentences, which is expressed by the first of the quoted axioms. The theory whose axioms say that the truth predicate is

[5]The notation $\mathsf{TB}\restriction$ is used more often in the literature.

[6]Obviously, we mean here the Gödel code representing the conjunction of these formulae. For the sake of clarity, we will use slightly imprecise expressions.

compositional is called CT^-[7]. The precise definition of this theory may be found in (Halbach, 2011). The other principle, which can be postulated, is the **extentionality principle** for the truth predicate:

> For any sentences $\varphi(t)$, $\varphi(s)$ from the language of the base theory, if the values of the terms t,s are equal, then the sentence $\varphi(t)$, is true if and only if the sentence $\varphi(s)$ is true.

Another possible requirement is that the sentences containing the truth predicate satisfy induction or, equivalently, the least number principle:

> Every nonempty subset of natural numbers defined with a formula containing the truth predicate has the least element.

In the language of first-order logic (in which all the theories considered here are formulated), the above principle can be expressed with an infinite system of axioms, the so called induction scheme for the formulae of the language extended with the truth predicate. The above principle has a more technical character than the ones which we have previously described. However, we can interpret it as follows: the properties defined using arithmetical predicates are "well defined" in the sense of not being vague.

Let us add that the theories CT^- and TB^- with the induction scheme for the sentences containing the truth predicate are called CT and TB, respectively. We hope that the Reader sees that there is a vast array of natural properties which the truth predicate should satisfy. There are even more possibilities, when we consider the self-referential truth predicate, that is, if we try to account for the behavior of the truth predicate applied to sentences in which that very predicate occurs.

1.2 Weak and strong truth theories

It is one of the very basic facts in the theory of truth that the theory CT proves certain arithmetical sentences which are not provable in PA alone. Namely, by Gödel's Second Theorem we know that if PA is consistent, then it does not prove the sentence $\mathsf{Con}_{\mathsf{PA}}$ which formalises the consistency claim for PA. However, the following fact holds:

Theorem 1 (Tarski)**.** CT *proves the sentence* $\mathsf{Con}_{\mathsf{PA}}$.

[7]Again, the theory in question is more often called $\mathsf{CT}\uparrow$. More generally, the theories which we denote Th^- are typically called $\mathsf{Th}\uparrow$.

Let us present an informal sketch of the proof of this theorem (a full proof may be found, for instance, in (Łełyk & Wcisło, 2017a)): We first show that CT proves the statement "All axioms of PA are true." Since PA has infinitely many axioms, this is not quite trivial. It is not enough to prove that every axiom separately is true (which can be done already in TB^-). We need to show the general statement. The intuition behind the proof is not terribly complicated, but it does contain some technical details, so we will only sketch it. Working in CT^-, let us fix any formula $\varphi(x)$. The sentence

$$T(\varphi(\overline{0})) \wedge \forall x(T(\varphi(\overline{x})) \rightarrow T(\varphi(\overline{x+1}))) \rightarrow \forall x T(\varphi(\overline{x}))$$

is an (actual) instance of the induction scheme (with parameter φ) for a certain formula with the predicate T, therefore it is available in CT as an axiom. Using the compositional axioms of CT^-, we obtain

$$T(\varphi(\overline{0}) \wedge \forall x(\varphi(x) \rightarrow \varphi(x+1)) \rightarrow \forall x \varphi(x))$$

and the above sentence states that the instance of the induction scheme for the formula φ is true. Since φ was arbitrary, we obtain the general sentence[8]. PA has only finitely many axioms except for the induction scheme, so by finitely many applications of compositionality of the truth predicate, we can show that all of them are true.

Having proved that the axioms of PA are true, we show by induction on the number of steps in a proof that any sentence which is derivable from the axioms of PA is true. At the same time, we can show that no sentence of the form $\varphi \wedge \neg\varphi$ is true. Therefore, no sentence of this shape is provable in PA which ends the sketch of the proof of Theorem 1.

The above theorem may be viewed philosophically important. It turns out that adjoining to PA a truth predicate satisfying very natural conditions yields a theory stronger than PA. This fact has been employed in a well-known argument against the deflationary theory of truth[9]. When Th_1, Th_2 are two theories such that $\mathsf{Th}_1 \subseteq \mathsf{Th}_2$ and there exists a sentence in the language of the theory Th_1 which is provable in Th_2, but not in Th_1, we say that Th_2 is **non-conservative** over Th_1. We say that Th_2 is conservative over Th_1 otherwise. By Theorem 1 (and Gödel's Theorem) it follows that CT is non-conservative over PA. Regardless of the philosophical importance of this

[8]We are omitting certain details here. For instance, the described argument only shows the truth of the parameter-free induction scheme. As we have said, the full proof requires us to deal with some technical issues which are not conceptually demanding.

[9]See (Ketland, 1999), (Shapiro, 1998).

specific fact, the following general question seems interesting: what properties of the truth predicate make the truth theory Th non-conservative over its base theory B? In the following paper, we describe certain results concerning this question. In other words, we try to understand what properties of the notion of truth make the truth theory "stronger" than its base theory.

2. Known results concerning conservativity

In light of Tarski's result discussed in the previous section that a compositional truth theory with full induction scheme for the whole language is not conservative over PA, it is natural to ask whether the truth theory CT^- is conservative, in which we assume only that the truth predicate is compositional. This is settled by the following theorem:

Theorem 2 (Kotlarski–Krajewski–Lachlan, Enayat–Visser, Leigh)**.** CT^- *is conservative over* PA*.*

Before we discuss the above theorem, let us comment upon its attribution. Kotlarski, Krajewski, and Lachlan (1981) proved a model-theoretic theorem which implied the conservativity of a certain theory very close to CT^-. When that paper was written, the axiomatic truth theories were not yet isolated as a separate field of research and their standard definitions were yet to be established. Therefore, the theory whose conservativity may be deduced from the Kotlarski–Krajewski–Lachlan's result is different from CT^- (it axiomatises satisfaction rather than truth) and it is not quite clear, how should we modify their proof in order to show the conservativity of CT^-. A conservativity proof of a compositional truth theory, much simpler than the argument in Kotlarski *et al.* (1981), has been obtained only by Enayat and Visser (2015). The theory which they investigated also was different from CT^-. This difference, however, was not so significant. The conservativity proof for CT^- has only been given by Leigh (2015) who used still other techniques.

2.1 Closure and Correctness Principles

Theorem 2 states that a purely compositional truth theory does not prove more arithmetical facts than PA alone. Only upon adding the induction scheme for the formulae containing the truth predicate will it allow us to prove, for instance, that PA is consistent.

Hence, we see two very natural theories CT^- and CT, only one of which is conservative. The induction for the truth predicate allows us to prove many facts about its structure. Compositionality, the basic feature of this predicate is not enough to prove new arithmetical theorems. Our question on the natural dividing line between truth theories which are conservative and not conservative over PA may be narrowed down to the following problem: what natural axioms characterising the truth predicate added to CT^- will make the resulting theory non-conservative over PA?

Ali Enayat suggested naming the dividing line between conservative and non-conservative truth theories between CT^- and CT **the Tarski boundary**[10]. Now our question may be expressed as follows: where is the Tarski boundary located? Let us discuss some natural axioms which extend CT^-, but are provable in CT. One very natural group of such axioms is the **closure and correctness principles**. Closure principles state that true sentences are closed under reasoning in a given deductive system. Correctness principles state that all sentences in a certain set are true. Let us present some principles of these sorts.

From the non-conservativity proof for CT, we may isolate one very simple correctness principle which definitely isn't conservaitve. Namely, **the principle of correctness of** PA:

> Every theorem of Peano arithmetic is true.

The above principle is also called **the global reflection principle** over PA. Let us notice that in the non-conservativeness proof for CT we have used exactly the fact that CT proves the correctness of PA.

We can isolate two further natural principles provable in CT which together imply the principle of correctness of PA. The first one is **the principle of closure under first-order logic**:

> Every sentence provable in first-order logic from true premises is true.

We can say that this principle is of more fundamental character than the principle of correctness of PA. It only says about the connection between

[10]Let us briefly justify the choice of the name. Tarski has been apparently the first one to point out the "weakness" of some arithmetic truth theories (inter alia TB^-). Besides that, CT^- is modelled after his inductive conditions defining the satisfaction relation. Ali Enayat and the authors of the paper has used this expression a few time in conference talks. However, to our best knowledge, this paper is the first place where the expression has been used in print.

truth and first-order logic and does not explicitly depend on our trust in the truth of the axioms of arithmetic. This trust is expressed by **the principle of axiomatic correctness of** PA:

Every axiom of PA is true.

Using standard proof-theoretic techniques, one can show that the arithmetical consequences of CT strictly contain the arithmetical consequences of CT^- with the principles of axiomatic correctness of PA and closure under first-order logic. Therefore, we have isolated a natural truth theory which is strictly weaker than CT but still not conservative over PA. It could seem that leaving any of these two axioms of this theory added to CT^- would also yield a non-conservative extension. However, it turns out that the principle of the axiomatic correctness of PA is one of the weak principles, which has been stated already in Kotlarski *et al.* (1981). These results have also been announced in (Enayat & Visser, 2015) and in (Leigh, 2015) (where it has been presented with proof). All of the cited sources bring different methods to demonstrate this theorem.

Theorem 3 (Kotlarski–Krajewski–Lachlan, Enayat–Visser, Leigh)**.** CT^- *with the principle of axiomatic correctness of* PA *is conservative over* PA*.*

Hence, it turns out that we can narrow down our search of the boundary between weak and strong compositional truth theories in a rather precise way. Adding to CT^- a principle that all axioms of PA are true is not enough to obtain new arithmetical consequences. On the other hand, extending this theory further with a principle that all sentences provable in PA are true, already turns out to be non-conservative.

2.2 Bounded induction scheme for the truth predicate

Another perspective which allows us to look for natural theories which are not conservative over Peano arithmetic but weaker than CT is restricting the induction scheme to some specific classes of formulae. We say that a formula is in the class Δ_0, if all quantifiers which appear in it are bounded, i.e., they are of the form $\forall x < t$, $\exists y < s$ for some terms t, s. Hence, the truth of sentences in the class depends only on objects of some fixed size. We can think of them as some special class of sentences whose truth value may be decided effectively. This class of formulae plays a very significant role in the research on metamathematical properties of arithmetic.

Another important class of formulae is Π_1. The formulae in this class are of the form

$$\forall x_1 \ldots \forall x_n \varphi,$$

where φ is a Δ_0 formula. We can think of them as purely universal formulae. They express that certain simple facts which may be decided effectively hold for all objects.

An important class of subtheories in Peano arithmetic are its fragments resulting from restricting the formulae in the induction scheme to the formulae of class Π_1 or Δ_0. We will follow this path also in the case of truth theories. By CT_1 we mean CT^- with all the instances of the induction scheme

$$\varphi(\overline{0}) \wedge \forall x(\varphi(x) \to \varphi(x+1)) \to \forall x \varphi(x)$$

in which the formula φ is an arbitrary formula of the class Π_1.

Let us note that the restriction to formulae in the class concerns only the formulae containing the truth predicate, since already CT^- contains all instances of the induction scheme for the arithmetical formulae, as an extension of PA.

The theory CT_1 is rather natural in the context of our research, since by inspection of the non-conservativity proof for CT, we can conclude that we in fact used only the axioms available in CT_1. We reach the following conclusion:

Theorem 4. CT_1 *is not conservative over* PA.

Indeed, one can easily show that the principle of correctness of PA, and consequently the principle of axiomatic correctness of PA, are provable in CT_1, as is the principle of closure under first-order logic. Therefore, we have reached another perspective allowing us to narrow down our search for the natural principles which make truth theory significantly stronger than its base theory. In particular, it is natural to ask about the conservativity of the theory CT_0 which results from restricting the induction scheme for formulae containing the truth predicate to Δ_0 formulae.

Let us add that TB (i.e., TB^- with the full induction scheme) is conservative over Peano arithmetic which is a significantly simpler result than Theorem 3. It follows that a truth theory with natural axioms including the full induction scheme does not automatically have to prove new arithmetical facts. Moreover, we can show examples of fully inductive theories nontrivially extending TB and based on some variants of Tarski's equivalences which are

still conservative. Therefore, the fact that we consider compositional truth theories is crucial for our results.

3. Discovering the Tarski Boundary

In this section we present the main known facts on the contour of the Tarski Boundary – most of the theorems, that we have stated, are yet unpublished, so the content of this section is to be treated as a report on a work in (hopefully) progress.

Let us start with briefly completing what we have already presented: in the last section we observed that (over CT^-) the closure under first-order logic principle in conjunction with the principle of axiomatic correctness of PA proves the principle of correctness of PA. It transpires that the first principle alone is capable of doing this: working in CT^- extended with the closure under the first-order logic principle, we will prove that each axiom of PA is true. The proof of this fact is rather standard and very intuitive: finitely many axioms fixing the basic rules of addition, multiplication and ordering are already true in CT^-. What remains are induction axioms: PA (even interpreted in a non-standard model) "thinks that" objects which it talks about are ordered as natural numbers, meaning that from 0 up to an element b there are only finitely many steps (obviously exactly b, which in a non-standard model can be a non-standard number). Working in CT^- with the closure under the first order logic principle and assuming $\varphi(\overline{0})$ and $\forall x(\varphi(x) \to \varphi(x+1))$ are true for a fixed arithmetical formula $\varphi(x)$, for an arbitrary a we will build a proof of $\varphi(\overline{a})$ in pure first order logic (we use the dictum de omni and modus ponens rules a many times). The proof runs in parallel to a classical argument that in the standard model of arithmetic the axioms of induction are true, with the only difference that in a non-standard model we use the induction axiom for a formula

$$\vartheta(x) := \mathsf{Prov}_{\mathsf{PA}}(\varphi(\overline{x})).$$

Since the set of true sentences is closed under reasoning in first-order logic, we can conclude that $\varphi(\overline{a})$ is true. Since a was arbitrary, we conclude that for all a $\varphi(\overline{a})$ is true, hence (on the basis of the compositional axioms) that the sentence $\forall x(\varphi(x))$ is true. This concludes our proof.

In an analogous way one can show (bypassing one additional difficulty – we refer the reader to (Cieśliński, 2010b) where this was proved for the first

time) that the principle of correctness of PA is equivalent to the following, much more restricted, correctness principle:

All validities of first-order logic are true.

The above principles might be naturally grouped into these claiming that the set of true sentences is closed under certain rules of inference (reasoning in first-order logic, for example) and these claiming that all sentences from a certain set are true (e.g. theorems of PA or theorems of first-order logic). Intuitively, the principles of the first kind say something more than their counterparts of the second kind. However, last year Cezary Cieśliński presented an insightful proof that over CT^- these principles are equivalent. Let us isolate it as separate

Theorem 5 (Cieśliński)**.** CT^- *extended with the axioms "All theorems of first-order logic are true" proves the closure under first-order logic principle.*

Searching for weaker principles provable in CT, but properly extending CT^- , let us extract from the closure under first-order logic **the principle of closure under propositional logic**:

> Each sentence provable in classical propositional calculus from true premises is true.

Obviously, over CT^- the above sentence is provable from the closure under first-order logic principle. As was shown by Cezary Cieśliński (2010a) CT^- , extended with the principle of closure under propositional logic, is equivalent to the compositional theory of truth with bounded induction, i.e. CT_0.

Theorem 6 (Cieśliński)**.** *The principle of closure under propositional logic is provable in* CT_0*. Each axiom of* CT_0 *is provable in* CT^- *extended with the principle of closure under propositional logic.*

Long before this paper of Cieśliński, Henryk Kotlarski (1986) published a proof that CT_0 proves the principle of correctness of PA. The argument, despite being concise, seemed correct and convincing enough to be cited also in Cieśliński (2010b) and Halbach (2011). However, in 2008 Albert Visser and Richard Heck noticed a gap in the proof of Koltarski: the problem was to demonstrate that CT_0 proves the sentence:

> Each axiom of first-order logic is true.

Kotlarski's proof worked fine for CT_0 extended with the above sentence. It is quite clear that the above can be proved with the help of the induction for Π_1 formulae. Reconstructing this proof with induction only for bounded formulae seemed so undoable that many logicians (including the authors of this paper) started searching for the proof that CT_0 lies on the conservative side of the Tarski Boundary.

Finally, it was shown (the proof appeared in (Łełyk & Wcisło, 2017b)) that CT_0 proves the same arithmetical sentences as CT_0 with the principle of correctness of PA added. Remarkably, in this proof only two very natural principles (provable in CT_0 but not in CT^-) were used: the first one was, introduced previously, the principle of axiomatic correctness of PA, the second was the generalized commutativity with the disjunction principle, called the disjunctive correctness principle:

> For all x and for every sequence of x sentences $\varphi_0, \ldots, \varphi_x$, their disjunction is true if and only if one of $\varphi_0, \ldots, \varphi_x$ is true.

Let us clarify this a little bit: for every natural number n, CT^-, using compositional axioms, will be able to prove that a disjunction of n sentences is true exactly when one of these sentence is. However, it will not be able to prove the above general statement[11]. It came as a surprise that such a simple generalization of compositional axioms, together with a (conservative when considered separately) principle of axiomatic correctness of PA, gives a theory which proves the same arithmetical sentences as CT^- with the PA correctness principle. Let us summarize this in the following:

Theorem 7 (W)**.** CT^- *extended with the principles of disjunctive correctness and the axiomatic correctness of* PA *proves the same arithmetical sentences as* CT^- *extended with the principle of correctness of* PA*.*

It is worth emphasizing that these results are not self-evident: so far it turns out that all the principles that we know to be located on the non-conservative side of the Tarski Boundary prove (at least) the consequences of the principle of correctness of PA (henceforth let us denote this principle with TPA). Let us observe that this set contains much more PA-unprovable sentences than simply the sentence naturally expressing the consistency of PA (abbreviated as $\mathsf{Con}_{\mathsf{PA}}$). We have already seen that this sentence is provable in $\mathsf{CT}^- + \mathsf{TPA}$. This is an arithmetical sentence, hence, applying

[11] This was first shown by Kotlarski *et al.* (1981). One can give an alternative proof based on Enayat and Visser methods of constructing full satisfaction classes.

finitely many times the compositional axioms, we can show that it is true. Since this theory proves the closure under first-order logic principle and we know that the axioms of $\mathsf{PA} + \mathsf{Con}_{\mathsf{PA}}$ are true, therefore no false sentence can be a consequence of this theory (in particular $0 = 1$ cannot). Thus we have just proved the consistency of theory $\mathsf{PA} + \mathsf{Con}_{\mathsf{PA}}$ i.e. the sentence:

$$\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}}}.$$

Nothing stops us from iterating this process further, this way proving stronger and stronger consistency assertions

$$\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}}}}$$

$$\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}}}}}$$

$$\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}}}}}}$$

and so on.

The arithmetical capacities of $\mathsf{CT}^- + \mathsf{TPA}$ does not stop there. It is not hard to convince oneself that it proves all sentences with the form

$$\forall x(\mathsf{Prov}_{\mathsf{PA}}(\varphi(x)) \to \varphi(x)) \tag{$*$}$$

for an arbitrary arithmetical formula $\varphi(x)$. The set of all sentences of this form is called the uniform reflection principle over PA[12]. A small subset of this set (for Π_1 formulae) is sufficient to prove all the above iterations of consistency statements. And that's not all: the set of (Gödel codes of) sentences of the above form is recursive (hence strongly representable in PA), hence in arithmetic we can define a theory

$$\mathsf{PA}^1 := \mathsf{PA} + \forall x(\mathsf{Prov}_{\mathsf{PA}}(\varphi(x)) \to \varphi(x))$$

for which the standard provability predicate will satisfy the Gödel–Löb conditions. In $\mathsf{CT}^- + \mathsf{TPA}$ we will prove all sentences of the form $(*)$ for PA^1, i.e. all sentences

$$\forall x(\mathsf{Prov}_{\mathsf{PA}^1}(\varphi(x)) \to \varphi(x)),$$

where $\varphi(x)$ ranges over arithmetical formulae with at most one free variable. In the next step we can define the theory PA^2, replacing in $(*)$ PA with PA^1. Iterating this process in the infinite, in the limit step taking

$$\mathsf{PA}^\omega := \bigcup_{n\in\omega} \mathsf{PA}^n$$

[12]It can be seen right now why we called the principle of correctness of PA the "global" reflection – in the presence of the truth predicate we can express the above principle in a single sentence.

we will obtain an arithmetical axiomatization of the arithmetical consequences of $\mathsf{CT}^- + \mathsf{TPA}$. A very elegant proof that PA^ω is really sufficient for deducing all arithmetical consequences of this theory of truth, was given by Henryk Kotlarski (1986).

The situation starts looking as if every "natural" theory of truth which proves the consistency of arithmetic, proved at the same time all the sentences from PA^ω. Obviously one can cook-up some artificial counterexamples to this "theorem": for example theory CT^- extended with the axiom "$\mathsf{Con}_{\mathsf{PA}}$ is true" is non-conservative over PA and much weaker than the considered "natural" theories (for example, it does not prove the sentence $\mathsf{Con}_{\mathsf{PA}+\mathsf{Con}_{\mathsf{PA}}}$). Obviously "natural" is not a formal notion, but it expresses a certain heuristics: it helps to temporarily block the ad hoc counterexamples. Right now we are trying to find a "natural" counterexample, possibly in the meantime realizing that no such counterexample can exist. Then we will probably understand what "natural" means.

Let us stress that in the above we did not say that CT_0 proves the principle of correctness of PA. The proof of non-conservativity of this theory consists in constructing a formula $T'(x)$ which, provably in CT_0 behaves like a predicate satisfying both CT_0 and the principle of correctness of PA. Using a definition the introduction of which we postpone for a moment (Definition 13), it has been shown that CT_0 augmented with the principle of correctness of PA is relatively truth definable in CT_0. However, we still didn't know whether the constructed formula $T'(x)$ *provably* in CT_0 had the same extension as the "original" truth predicate. Stating this less formally: it could be the case that CT_0 is able to "upgrade" its own truth predicate but cannot prove that its own truth predicate is as good (i.e. satisfies the principle of correctness of PA). In the meantime Ali Enayat[13] showed that focusing on the extension of CT^- with the principles of the axiomatic correctness of PA and disjunctive correctness, we were not in fact working with a weaker theory. He proved the following:

Theorem 8 (Enayat)**.** CT^- *extended with the principle of axiomatic correctness of* PA *and the disjunctive correctness principle proves* CT_0*.*

It turned out that, up to deductive equivalence and looking only at the theories that we can prove to be non-conservative, there are only two minimal theories above the Tarski Boundary: CT_0 and $\mathsf{CT}^- + \mathsf{TPA}$, and, moreover, that they are mutually relatively truth definable (Definition 13).

[13]Personal communication.

After all, it has been shown that this picture is even simpler: a direct fix to the old proof of Kotlarski was discovered. Let us summarize our findings in the following:

Theorem 9 (Cieślinski, Enayat, Kotlarski, Ł)**.** *The following theories are equivalent:*

1. CT_0.
2. CT^- *with the principle of correctness of* PA.
3. CT^- *with the principle of closure under first-order logic.*
4. CT^- *with the principle of closure under propositional logic.*
5. CT^- *with the principle of correctness of first-order logic.*
6. CT^- *with the principles of disjunctive correctness and axiomatic correctness of* PA.

So far the situation on the Tarski Boundary looks as if there were the least ("natural") theory which admits many different axiomatizations. Obviously, as the careful Reader has certainly noticed, some questions have been left unanswered in the above considerations. This was not accidental: as for this moment we still do not know whether the extension of CT^- only with the principle of disjunctive correctness is conservative over PA or not[14]. Intuitively, it should be a weak extension of PA. However it would not be the first time when our intuitions have failed...

It is worth mentioning, at least concisely, the possible impact of the above theorem on the philosophical debate over deflationism[15]. Assuming that the deflationist should present a theory which both proves some general facts about the truth predicate and is conservative over PA, it follows that his options are rather limited. He cannot, for example, demand both classical compositionality and closure under propositional logic from the truth predicate axiomatized by such a theory. Furthermore, he cannot even demand generalized compositionality (which would imply the disjunctive correctness principle) and the principle of correctness of PA. It might seem that this

[14] Telling the truth: we did not know this when writing the polish version of this paper. Recently, however, Fedor Pakhomov presented an insightful proof that CT^- with the principle of disjunctive correctness is actually the same theory as CT_0. This is a highly unexpected result.

[15] We thank the anonymous referee for the suggestion of adding this remark to the paper.

situation is hopeless. To obtain this conclusion, however, we need to assume that the deflationary theory of truth needs to prove general facts about the behavior of the truth predicate and be conservative. This requirement is based on yet another assumption: we have to agree that provability in a theory is a good enough explication of the notion of justification or explanation (depending on how the thesis of deflationism is formulated). This view has been recently ciriticised at length in (Cieśliński, 2017) and we have to admit that right now we are unconvinced as to whether Theorem 9 can really play a role in this debate.

3.1 The Tarski Boundary and different truth theories

Let us observe that we can also ask about the contour of the Tarski Boundary with respect to theories of truth different from CT^-. For example, we can start from the least (thus far) "natural" non-conservative theory of truth i.e. CT_0 and weaken the compositional axioms, modelling them not after the classical logic, but, for example, on strong Kleene logic. In such a theory, known as PT_0[16] we do not have a global axiom for the negation, i.e.

> (For every sentence φ) The negation of φ is true if and only if φ is not true.

Instead for every connective (negation included) and quantifier we say separately when the negation of a sentence beginning with this connective is true. For example, the following sentence is an axiom of PT_0

> (For all arithmetical sentences φ, ψ) The negation of the conjunction of φ and ψ is true if and only if the negation of either of φ or ψ is true.

We can now ask: does the contour of the Tarski Boundary depend on which logic we choose for the compositional and Δ_0 inductive truth predicate? This question is one of the topics of our current research[17].

[16] More precisely, in the literature only the non-inductive version of this theory, denoted PT^- (or $\mathsf{PT}\restriction$) is known, but PT_0 is simply this theory with axioms of induction for the Δ_0 formulae of the extended language.

[17] Now we know that some extensions of PT^- (compare footnote 16) are strong but weaker than CT_0. The question whether every natural strong extension of CT^- proves Global Reflection is still open.

4. Other Notions of Conservativity

The question about the conservativity of a given theory is just the first step into differentiating various axiomatic theories of truth. It can be taken as the first approximation, classifying the theories as either strong or weak. More generally we can ask which theories are stronger than other theories (in particular: comparing the non-conservative theories). In this the following (obvious in fact) generalization of the notion of conservativity can be used:

Definition 10. A theory Th_1 is syntactically stronger than Th_2 if and only if the arithmetical consequences of Th_2 form a proper subset of the set of arithmetical consequences of Th_1.

One can show, for example, that CT_1 is syntactically stronger than CT_0 and CT is stronger than CT_1. Non-stratified, compositional and fully inductive theories of truth are usually still much stronger, for example FS is stronger than CT, KF than FS and VF than KF[18].

Observe, however, that distinguishing theories only on the basis of their arithmetical consequences blurs the differences between many theories, whose axioms have intuitively a very different character. For example, both CT^- and TB are syntactically conservative over PA, hence they cannot be told apart solely on the base of their arithmetical consequences. One can consider also a different measure which would enable us to differentiate between theories with the same syntactical strength. This measure is based on a, well-known from the literature, notion of semantical conservativity:

Definition 11. A theory Th is semantically conservative over PA if and only if every model of PA can be expanded (with preservation of the universe and arithmetical functions) to a model of Th.

The philosophical intuition motivating this notion is as follows: we think about models of a theory as "possible worlds" ("possible" from the point of view of the considered theory). If a model of PA cannot be extended to a model of Th, it means that such possibility, while admitted by PA, is excluded by Th. It is worth noticing that semantical conservativity implies syntactical conservativity (by Completeness Theorem), but it does not reverse: neither TB, nor CT^- is semantically conservative. This notion can be generalized in the following way:

[18]These names are standard in the literature. Definitions of KF and FS can be found in (Halbach, 2011), whereas VF was introduced in Cantini's paper (1990).

Definition 12. A theory Th_1 is semantically stronger than a theory Th_2 if and only if the class of models of PA that can be expanded to models of Th_1 is a proper subclass of the class of models of PA, that can be expanded to models of Th_2.

Basing on the intuition just introduced, we can say that Th_1 is semantically stronger than Th_2, if Th_1 eliminates more "possible worlds", than Th_2. Using this distinction one can prove that TB is semantically weaker than CT^-, which matches our intuitions that compositional axioms "say more" about the notion of truth, than Tarski biconditionals (even when considered in the presence of full induction).

The most fine-grained relation which can differentiate between various truth theories, was introduced by Kentaro Fujimoto in (2010) and is known as relative truth definability.

Definition 13. Let Th_1 and Th_2 be two truth theories. We say that Th_1 is relatively truth definable in Th_2 if and only if there exists a formula $\varphi(x)$ such that Th_2 proves the axioms of Th_1 with $\varphi(x)$ substituted for the truth predicate of Th_1.

To say the same things in simple words (perhaps less precisely): Th_1 is relatively truth definable in Th_2 if Th_2 can define the truth predicate which satisfies the axioms of Th_1. We shall say that Th_2 is Fujimoto-stronger than Th_1 if Th_1 is relatively truth definable in Th_2 but not vice-versa. The proof of Theorem 7 shows that CT_0 together with the principle of correctness of PA is relatively truth definable (hence not Fujimoto-stronger than) in CT^- with the principles of disjunctive correctness and axiomatic correctness of PA. There are theories which can be distinguished only by the above relation, for example TB^- and UTB^-[19].

5. Summary and Open Problems

We began the paper with introducing the most basic measure of strength of axiomatic theories of truth, according to which a theory is classified as strong if it proves some sentences which are unprovable in PA. The boundary between strong and weak axiomatic theories of truth was called the Tarski Boundary. The most important discovery concerning the contour of the Tarski Boundary can be summarized as follows: each "natural" theory of

[19]Defined in Halbach (2011) as $\mathsf{TB}\restriction$ and $\mathsf{UTB}\restriction$.

truth, which up till now has proved to be strong, proves CT^- with the principle of correctness of PA[20]. Moreover this last theory admits many different axiomatizations, one of them being CT^- augmented with a scheme of induction for bounded formulae with the truth predicate (this theory was called CT_0). Lastly, we showed that there exist two interesting strengthenings of the introduced measure, which help us to discern between the strength of truth theories for which the basic notion was too coarse-grained. It is worth emphasizing that still there are many interesting open questions concerning the "strength" of axiomatic truth theories. We list some of them below:

1. Is CT^- with the principle of disjunctive correctness conservative over PA[21].

2. Is CT^- with the principle of the correctness of propositional logic conservative over PA? Let us notice that the above additional axiom is a correctness principle corresponding to the principle of closure under propositional logic. The latter one (over CT^-) is equivalent e.g. to the Global Reflection principle, hence is very strong.

3. Is CT^- semantically stronger than UTB? We know that $\mathsf{CT}^-$is at least as strong as UTB (i.e. every model which can be expanded to a model of CT^-, can be expanded to a model of UTB; it was proved in Łełyk and Wcisło 2017a).

4. Does CT^- relatively truth define UTB?[22]

Bibliography

Cantini, Andrea (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic*, *55*(1), 244–259.

Cieśliński, Cezary (2017). *The epistemic lightness of truth. Deflationism and its logic*. Cambridge: Cambridge University Press.

[20] Actualization: now we know quite a few such theories, one of them being PT_0, compare footnote 17. The question is still open for extensions of CT^-.

[21] Actualization: this question was already answered by Fedor Pakhomov, compare footnote 14.

[22] Recently, this has been answered in the negative by Albert Visser (private communication).

Cieśliński, Cezary (2010a). Deflationary truth and pathologies. *Journal of Philosophical Logic*, *39*(3), 325–337.

Cieśliński, Cezary (2010b). Truth, conservativeness, and provability. *Mind*, *119*(474), 409–422.

Enayat, Ali, Visser, Albert (2015). New constructions of satisfaction classes. In Theodora Achourioti, Henri Galinon, José Martinez Fernández (Eds.), *Unifying the philosophy of truth* (pp. 321–335). Dordrecht: Springer Netherlands.

Franzen, Torkel (2003). *Inexhaustibility. A Non-Exhaustive Treatment*. Association for Symbolic Logic. Wellesley: A K Peters.

Fujimoto, Kentaro (2010). Relative truth definability of axiomatic truth theories. *Bulletin of Symbolic Logic*, *16*(3), 305–344.

Halbach, Volker (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Kaufmann, Matt, Schmerl, James (1987). Remarks on weak notions of saturation in models of Peano Arithmetic. *Journal of Symbolic Logic*, *52*(1), 129–148.

Kaye, Richard (1991). *Models of Peano Arithmetic*. New York: Clarendon Press.

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

Kotlarski, Henryk (1986). Bounded induction and satisfaction classes. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, *32*(31–34), 531–544.

Kotlarski, Henryk, Krajewski, Stanisław, Lachlan, Alistair (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, *24*, 283–293.

Leigh, Graham (2015). Conservativity for theories of compositional truth via cut elimination. *Journal of Symbolic Logic*, *80*(3), 845–865.

Łełyk, Mateusz, Wcisło, Bartosz (2017a). Models of weak theories of truth. *Archive for Mathematical Logic*, *56*(5–6), 453–474.

Łełyk, Mateusz, Wcisło, Bartosz (2017b). Notes on bounded induction for the compositional truth predicate. *The Review of Symbolic Logic*, *10*(3), 455–480.

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

Originally published in *Studia Semiotyczne* 31/1 (2017), 23–44.