STUDIA SEMIOTYCZNE, t. XXXIII, nr 2 (2019), s. 255–286
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiii2.08

Article

MATEUSZ ŁEŁYK [*]

COMPARING AXIOMATIC THEORIES OF TRUTH [1, 2]

SUMMARY: This paper is a direct continuation of (Łełyk, Wcisło, 2017), where one particular way of comparing axiomatic theories of truth was discussed at length in the context of extensions of one concrete truth theory—the basic stratified theory of compositional truth, known under the name CT$^-$. Other possible truth theories and ways of comparing them were only briefly sketched. In the current study we want to describe this field of research in greater details. In Section 2 we shall introduce various axiomatic theories of truth. The formal definitions will be supported by intuitive explanations. In Section 3 we introduce three relations that can serve as explications of strength of a theory. One of them will be, known from (Łełyk, Wcisło, 2017) the proof-theoretical strength, and two others, model-theoretical strength and Fujimoto definability, will be its refinements. In the last section we shall discuss the formal results showing how these relations shape the universe of theories under investigation.

KEYWORDS: axiomatic theories of truth, conservativity, Peano Arithmetic, disquotation, Kripke-Feferman, Friedman-Sheard.

1. INTRODUCTION

Suppose that we are working with a formal theory Th formulated in a language $\mathcal{L}$, which is capable of expressing syntactical properties of formulae

[*] University of Warsaw, Faculty of Philosophy and Sociology. E-mail: mlelyk@uw.edu.pl. ORCID: 0000-0001-5286-4511.

[1] The research presented in this paper was supported by The National Science Centre, Poland (NCN), grant number 2017/27/B/HS1/01830.

[2] Supported by the Foundation for Polish Science (FNP).

of $\mathcal{L}$. By this we mean, as a first approximation, that there exists a procedure of assigning names to syntactical terms and formulae of $\mathcal{L}$, i.e. a function which given any term t or a formula φ returns a closed term of $\mathcal{L}$, which we denote $\ulcorner t \urcorner$ or $\ulcorner \varphi \urcorner$ respectively. $\ulcorner \varphi \urcorner$ is to be thought of as φ taken into quotation marks. Secondly, Th must be able to correctly describe syntactical relations between various φ's. By this we mean, to give some examples, that there exist formulae in the language of Th Conj(x, y, z) (to express "x is a conjunction of y and z") and Subst(x, y, z, w) (to express "x results from y by substituting name z for variable w") such that, for example, the following are provable in Th:

$$\forall y, z\, (\mathrm{Conj}(\ulcorner \varphi \wedge \psi \urcorner, y, z) \rightarrow\ y = \varphi \wedge \mathrm{z} = \psi$$

$$\forall x\, (\mathrm{Subst}(x, \ulcorner y = y \urcorner, \ulcorner 0 \urcorner, \ulcorner y \urcorner) \rightarrow\ x = \ulcorner 0 = 0 \urcorner).$$

The first one expresses within Th that the only candidates for conjuncts of $\varphi \wedge \psi$ are φ and ψ and the second one that only 0 = 0 is the result of substitution of 0 (as a symbol) for variable x in formula $x = x$. In formal terms, we demand that all the primitive recursive functions be strongly represented in Th, but the Reader may safely use the intuitive description above.

It can be shown that already very modest arithmetical theories meet the requirements put forward in the above paragraph. Hence already very basic theories can talk about their own syntax. In striking contrast to the above, no such theory is capable of expressing the notion of truth for its own language.

Theorem 1 (Tarski). *Let* Th *be a consistent theory which is sufficiently strong to develop basic theory of syntax. Then there is no formula $T(x)$ such that for every sentence ψ in the language of* Th, Th *proves*

$$T(\ulcorner \psi \urcorner) \equiv \psi. \qquad (\mathrm{TB}(\psi))$$

It is a very basic intuition that every theory which is to be called a truth theory for Th should deliver all (i.e. for every ψ in the language of Th) the equivalences (TB(ψ)). To the group of such sentences we shall refer also as a T-scheme. In the effect the theorem tells us that in order to develop the theory of truth for Th we have to (at least) enrich its vocabulary. The most straightforward step now is simply to add the predicate for the property "x is a true sentence" along with some axioms governing its use. And this is precisely how an axiomatic theory of truth is born:

Definition 2. We say that P is an axiomatic theory of truth over Th if P is formulated in the language of Th extended with a fresh unary predicate T and for every ψ in the language of Th, (TB(ψ)) is provable in P.

In the next section we show that theories satisfying the above definition are very diverse.

2. A Variety of Truth Theories

Let us start by drawing two most basic classification lines between axiomatic theories of truth. They concern rather the syntactical shape of the chosen axioms for the truth predicate and serve to illustrate approaches that one can take in designing a theory of truth.

We say that a theory of truth P is typed (or: stratified) if the axioms of P regulate the behaviour of the truth predicate T only on sentences without the predicate T. Otherwise, P is untyped. We say that P is disquotational if P is axiomatized by sentences of the form (TB(ψ)), for various ψ. We say that P is compositional if it is axiomatized by (in addition to the axioms of Th) finitely many sentences determining how the truth of a formula depends on the truth of its subformulae.

Let us now illustrate this distinction with some canonical examples. For this purpose, we focus on one concrete choice of the base theory Th, and assume that it is always Peano Arithmetic (PA). It is given by finitely many axioms stating basic properties of addition, multiplication and successor (denoted with S) functions, such as

$$\forall x, y \; x \cdot S(y) = x \cdot y + x$$

and infinitely many axioms of induction of the form

$$\psi(0) \wedge \forall y (\psi(y) \rightarrow \psi(y+1)) \rightarrow \forall y \psi(y) \qquad (\mathrm{Ind}(\psi))$$

where ψ is in the language of arithmetic and may contain free variables other than y. The above sentence expresses the principle of induction for the set defined by $\psi(y)$: if 0 is in this set and if $y + 1$ is in it whenever y is, for arbitrary number y, then every number is its member. We shall use $\mathcal{L}$ to denote the language of arithmetic and $\mathcal{L}_T$ to denote $\mathcal{L}$ enriched with a fresh predicate T. The choice of PA as a base theory was commented already in the earlier paper (Łełyk, Wcisło, 2017).

Each theory P presented below results as an answer to the following questions:

1. Should P be typed or untyped?
2. Should P be disquotational or compositional?
3. For which formulae ψ of $\mathcal{L}_T$ should we add to P the axiom of induction (Ind(ψ))?

Question 3 is normally answered by allowing axioms (Ind(ψ)) for ψ from a particular level of the arithmetical Σ_n hierarchy. The basic Σ_0 (which, by definition, equals Π_0) level contains only formulae ψ in which every quantifier is bounded, i.e. whenever a formula of the form $\exists x\varphi$ occurs in ψ, then φ is of the form $(x < t \wedge \varphi')$, where t is a term that does not contain variable x (dually for the universal quantifier). We say that ψ is in the class Π_{n+1} if ψ is of the form $\forall x_1 \ldots \forall x_k \varphi$, where $\varphi \in \Sigma_n$. Dually, $\psi \in \Sigma_{n+1}$ if $\psi = \exists x_1 \ldots \exists x_k \varphi$ where $\varphi \in \Pi_n$.

2.1. Typed and Disquotational Theories

These can be considered the simplest truth theories as they are axiomatized solely by (variations of) the *T*-scheme. The most basic such theory is called TB$^-$:

Definition 3. TB$^-$ is PA together with all sentences TB(ψ) for ψ—a sentence in $\mathcal{L}$. If we allow induction axiom (Ind(ψ)) for every $\psi \in \mathcal{L}_T$, then the respective theory is denoted with TB.

TB$^-$ correctly guesses the truth of sentences. For example,

$$\mathrm{TB}^- \vdash T(\ulcorner 0 = 0 \urcorner).$$

Indeed the above holds, since $T(\ulcorner 0 = 0 \urcorner) \equiv 0 = 0$ is an axiom of TB$^-$ and $0 = 0$ is provable in PA. Similarly, if ψ is any arithmetical formula, then the (name of the) axiom of induction for ψ is provably within the scope of *T*.

However, even in the presence of full induction, the TB scheme alone has very strong limitations. For example it does not prove that every sentence of the form $t = t$ is true, where t is an arbitrary closed term. Formally

$$\mathrm{TB}^- \nvdash \forall t \;\; T(\ulcorner t = t \urcorner). \tag{1}$$

The above requires some explanation: for every concrete term t the sentence $T(\ulcorner t = t \urcorner)$ will be provable in TB$^-$, exactly by the reasons given above. However using induction in PA we can prove the universal sentence, that for every term t there exists a formula which results by the substitution of t in $x = x$ for every free occurrence of variable x. $\forall t \;\; T(\ulcorner t = t \urcorner)$ abbreviates that every such sentence is true. For the proof of 1 see (Halbach, 2011).

To overcome this obstacle one can extend the TB scheme, introducing uniform Tarski biconditionals. To introduce them we need one more arithmetical function: the value of a term. Let us recall that for any arithmetical term we can define its value (i.e. the natural number denoted by it) by the following recursion:

$$0^\circ = 0$$
$$S(s)^\circ = S(s^\circ)$$
$$(s + t)^\circ = s^\circ + t^\circ$$
$$(s \cdot t)^\circ = s^\circ \cdot t^\circ$$

Now such a definition can be formalized in PA, hence there exists an arithmetical formula $y = x^\circ$ such that

$$\mathrm{PA} \vdash \forall t \exists! y\ y = t^\circ.$$

For the sake of readability, we shall treat the symbol $\cdot^\circ$ as it was actually a term-forming expression. For every formula $\psi(x_1,\ldots,x_n)$ we define UTB(ψ) to be the sentence

$$\forall t_1 \ldots \forall t_n\ T(\ulcorner \psi(t_1,\ldots,\ t_n)\urcorner) \equiv \psi(t_1{}^\circ,\ldots,\ t_n{}^\circ). \qquad (\mathrm{UTB}(\psi))$$

For every concrete $\psi(x_1,\ldots,x_n)$ the above sentence says that for all terms $t_1,\ldots,$ t_n the sentence resulting by substituting t_i for x_i, for $i \leq n$, in $\psi(x_1,\ldots,x_n)$ is true if and only if $\psi(a_1,\ldots,a_n)$ holds where for each i, a_i is the value of term t_i.

Definition 4. UTB$^-$ extends PA with all sentences of the form UTB(ψ) for ψ—an arithmetical formula. UTB extends PA with induction axioms for all formulae of $\mathcal{L}_T$.

With this definition it is a straightforward observation that for every arithmetical formula $\psi(x)$

$$\mathrm{UTB}^- \vdash T(\ulcorner \forall x\ \psi(x)\urcorner) \equiv \forall t\ T(\ulcorner \psi(t)\urcorner),$$

i.e. the universal sentence $\forall x\ \psi(x)$ is true if and only if $\psi(t)$ is true for every term t (in the sense of PA). In particular,

$$\mathrm{UTB}^- \vdash \forall t\ T(\ulcorner t = t\urcorner).$$

However, there are still very important limitations to uniform Tarski biconditionals: the theory does not have the resources needed to establish the universal validity of even the simplest propositional schemes. For example

$$\mathrm{UTB} \nvdash \forall \varphi\ T(\ulcorner \varphi \vee \neg\varphi \urcorner),$$

where $T(\ulcorner \varphi \vee \neg\varphi \urcorner)$ abbreviates: “the disjunction of φ and $\neg\varphi$ is true”. For the proof of it, once again consult (Halbach, 2011).

Remark 5 (Historical remark). The first results about typed disquotational theories of truth were obtained already by Tarski in (1995), who immediately noticed their limitations.

2.2. Typed Compositional Theories

The limitation demonstrated at the end of the previous section points to the fact that we cannot write the axiom for the truth predicate T in the form

$$\forall\varphi\ T(\varphi) \equiv \varphi,$$

since, if φ is quantified, then it should occupy the place of a variable in the formula within the range of the quantifier. The way out of the problem is to choose axioms which would give us a procedure of determining the truth value of an arbitrary $\mathcal{L}$ sentence. In particular we can determine how the truth of a complex sentence depends on the truth of its immediate constituents. We give two examples of such sets of axioms and start from the less obvious one:

Definition 6. PT^{-} extends PA with the following finitely many axioms for the truth predicate:

1. $\forall s, t\ [T(\ulcorner s = t\urcorner) \equiv s^{\circ} = t^{\circ}]$.
2. $\forall s, t\ [T(\ulcorner \neg s = t\urcorner) \equiv s^{\circ} \neq t^{\circ}]$.
3. $\forall\varphi\ [T(\ulcorner\neg\neg\varphi\urcorner) \equiv T(\varphi)]$.
4. $\forall\varphi, \psi\ [T(\ulcorner \mathrm{\varphi} \wedge \psi\urcorner) \equiv T(\varphi) \wedge T(\psi)]$.
5. $\forall\varphi, \psi\ [\mathrm{T}\ (\ulcorner\neg(\varphi\wedge\psi)\urcorner) \equiv T(\ulcorner\neg\varphi\urcorner) \vee T(\ulcorner\neg\ \psi\urcorner)]$.
6. $\forall v \in \mathrm{Var} \forall\varphi \in \mathrm{Form}^{v}\ \ [T(\ulcorner\forall v\varphi\urcorner) \equiv \forall t\ T(\varphi[t/v])]$.
7. $\forall v \in \mathrm{Var} \forall\varphi \in \mathrm{Form}^{v}\ \ [T(\ulcorner\neg\forall v\varphi\urcorner) \equiv \exists t\ T(\ulcorner\neg\varphi[t/v]\urcorner)]$.
8. $\forall t, s\ \forall v \in \mathrm{Var} \forall\varphi \in \mathrm{Form}^{v}\ [s^{\circ} = t^{\circ} \rightarrow T(\varphi[s/v]) \equiv T(\varphi[t/v])]$.

For a natural number n, PT$_n$ denotes the result of adding to PT^{-} the axioms of induction for Σ_n formulae with the truth predicate, i.e. all sentences Ind(ψ) for ψ—a Σ_n formula of $\mathcal{L}_T$.

Let us have a word of comment about the meaning of the above axioms: by writing $\forall\varphi$ we implicitly quantify over (codes of) $\mathcal{L}$ sentences (not $\mathcal{L}_T$, though!) and the intended reading of, e.g., axiom 3 is: “For every arithmetical sentence φ, the sentence resulting from φ by adding two negations at the front of φ is true if and only if φ is true.”

In axiom 6 Var denotes the set of variables (more precisely: a formula “x is a variable”) and $\forall v \in$ Var is to be read as “For every variable v…”. $\forall\varphi \in$ Form^{v} expresses “For all formulae φ with at most one free variable v”. Consequently, the whole axiom 6 reads: “For every variable v and for every formula φ with at most v free, the sentence $\forall v\varphi$ is true if and only if for every term t the sentence resulting from φ by substituting term t for every free occurrence of variable v is true.”

The above axiom is correct, since for every natural number n there exists a term naming it, for example

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

and this procedure formalizes in PA (such terms are defined by a routine induction), hence

$$\mathrm{PA} \vdash \forall x \exists t \ \ t^{\circ} = x.$$

Terms of the form (Num) are called canonical numerals. The canonical numeral for n will be denoted $\underline{n}$. Obviously, there are more terms than such ones. Axiom 7 states that as long as two terms denote the same number it is irrelevant for the truth of a formula, which of them is used.

The Reader may wonder what does "P" in PT stands for and why we build a compositional theory of truth for arithmetical sentences without including the axiom

$$\forall \varphi \in \mathrm{Sent}_{\mathcal{L}} \ (T(\neg\varphi) \equiv \neg T(\varphi)). \qquad (\mathrm{NEG}(\mathcal{L}))$$

on the list. The answer to both questions is the same: we want the truth predicate T to appear only positively, i.e. non-negated, in the truth axioms. This makes such a theory easier to investigate and allows to study questions about how much (NEG($\mathcal{L}$)) contributes to the properties of axiomatic theories. Moreover, there is a very intuitive picture in the background of the positive aspect of PT^{-}: we think of the truth of sentences as being evaluated in stages. At the first stage atomic sentences are classified as either true, i.e. $T(\ulcorner s = t \urcorner)$ holds, or false i.e. $T(\ulcorner \neg(s = t) \urcorner)$ holds. For example after this stage we will be able to conclude that $T(\ulcorner 0 = 0 \urcorner)$ and $T(\ulcorner \neg 0 = S(0) \urcorner)$ hold. At the next stage, and more generally at any later stage, sentences of the form $\neg\neg\varphi$, $(\varphi \wedge \psi)$, $\neg(\varphi \wedge \psi)$, $\forall x \varphi$ and $\neg\forall x \varphi$ get evaluated for all φ, ψ which were evaluated at an earlier stage.[3] A sentence φ is true if at some stage φ is classified as such. It is false if $\neg\varphi$ is classified as true. In such a way if φ is classified as true or false, then we are given a procedure which reduces the problem of φ's truth value to the problem of truth value for atomic sentences. In such a situation we say that φ is grounded.

Observe that the axiom (NEG($\mathcal{L}$)) does not quite match this picture: we should not classify $\neg\varphi$ as true simply because φ was not classified as true at an earlier stage. This is because φ might require a longer procedure of evaluation and finally it will be classified as true. For example, after the first stage the sentence

[3] There is one subtlety in here: according to the axiom 5 it is sufficient for one of φ, ψ to get evaluated as false to classify $\varphi \wedge \psi$ as false.

$$(0 = 0) \wedge (0 = 0),$$

is not classified as true. However it would be a mistake to classify $\neg((0 = 0) \wedge (0 = 0))$ as true based on this.

Since for every sentence φ of $\mathcal{L}$ we can build such a procedure of evaluation (this is shown by induction on the complexity of φ), it can easily be shown that PT$^-$ is indeed a theory of truth. In fact, it properly extends UTB$^-$:

Proposition 7. *For every* $\varphi(x_1,\ldots, x_n)$,

$$\mathrm{PT}^- \vdash \forall t_1,\ldots,\ t_n\ T(\ulcorner\varphi(t_1, \ldots , t_n)\urcorner) \equiv \varphi(t_1{}^\circ,\ldots,\ t_n{}^\circ).$$

We urge the Reader that this does not show that (NEG($\mathcal{L}$)) is provable in PT$^-$, for the above proposition shows only that for every φ separately

$$\mathrm{PT}^- \vdash T(\ulcorner\neg\varphi\urcorner) \equiv \neg T(\ulcorner\varphi\urcorner).$$

However, PT$^-$ lacks induction axioms for formulae with the predicate T, so it cannot internalize the reasoning in the proof of Proposition 7. Upon adding the axiom (NEG($\mathcal{L}$)), PT$^-$ axioms 2., 3., 5. and 7. become redundant. The resulting theory is called CT$^-$:

Definition 8. CT$^-$ extends PA with the following axioms for the truth predicate:

1. $\forall s, t\ [T(\ulcorner s = t\urcorner) \equiv s^\circ = t^\circ]$.
2. $\forall \varphi\ [T(\ulcorner\neg\varphi\urcorner) \equiv \neg T(\varphi)]$.
3. $\forall \varphi, \psi\ [T(\ulcorner\varphi\wedge\psi\urcorner) \equiv T(\varphi)\wedge T(\psi)]$.
4. $\forall v \in \mathrm{Var}\forall\varphi \in \mathrm{Form}^v\ \ [T(\ulcorner\forall v\varphi\urcorner) \equiv \forall t\ T(\varphi[t/v])]$.
5. $\forall t, s\ \forall v \in \mathrm{Var}\forall\varphi \in \mathrm{Form}^v\ [s^\circ = t^\circ \rightarrow T(\varphi[s/v]) \equiv T(\varphi[t/v])]$.

As usual, CT$_n$ denotes CT$^-$ extended with Σ_n induction for formulae of $\mathcal{L}_T$ and CT admits a full $\mathcal{L}_T$ induction scheme.

The above theory should look familiar, as its axioms are straightforward formalisations of inductive Tarski truth conditions in $\mathcal{L}_T$. Let us point out some properties of CT$^-$:

Proposition 9. CT$^-$ $\vdash$ PT$^-$.

Sketch of the proof. Let us show how to prove the axiom 7. We work in CT$^-$. Fix a variable v and a formula $\varphi(v)$ with at most one free variable as shown. By (NEG($\mathcal{L}$)) $T(\ulcorner\neg\forall x\varphi\urcorner)$ is equivalent to $\neg T(\ulcorner\forall x\varphi\urcorner)$. This, in turn, by axiom 4. of CT$^-$ is equivalent to $\neg\forall t T(\varphi[t/v])$. The last formula, by De Morgan laws for quantifiers, is equivalent to $\exists t \neg T(\varphi[t/v])$. Since for every t, $\varphi[t/v]$ is a sentence the above is equivalent to $\exists t T(\ulcorner \neg\varphi[t/v]\urcorner)$, which ends the proof of axiom 7. □

Similarly, in CT$^-$ we can show that the law of non-contradiction (to give an example) holds for all formulae, i.e.

Proposition 10. CT$^- \vdash \forall\varphi\ T(\ulcorner\neg(\varphi\wedge\neg\varphi)\urcorner)$.

Proof. Work in CT$^-$ and fix φ. Observe that $\neg(T(\varphi)\wedge\neg T(\varphi))$ holds, since this is an instantiation of logical tautology. By the subsequent use of the axiom for negation, conjunction and once more negation we show that $\neg(T(\varphi)\wedge T(\ulcorner\neg\varphi\urcorner))$, $\neg T(\ulcorner(\varphi\wedge\neg\varphi)\urcorner)$ and finally $T(\ulcorner\neg(\varphi\wedge\neg\varphi)\urcorner)$ hold. □

Later on (in Section 2.5) we shall demonstrate some natural, but unprovable in CT$^-$ properties of the truth predicate. Let us now turn to untyped compositional theories.

2.3. Untyped Compositional Theories

As famously argued by Kripke (see Kripke, 1975), typing seems to be a way too radical response to the Liar paradox. There are many sentences with the truth predicate whose truth conditions are unproblematic, for example

“‘0 = 0’ is true.”

is true while “‘0 = 1’ is true and ‘0 = 0’ is true” is false. In this section we introduce axiomatic theories which are capable of dealing with such examples correctly.

What does it mean that a theory is compositional? We certainly expect such a theory to state the truth conditions for atomic sentences and then to characterize, that way or another, the truth conditions for complex sentences in terms of the truth conditions of its immediate subformulae. If we want our theory with the truth predicate T to correctly characterize the truth of all atomic sentences, then it should prove an analogon of axiom 1. of CT$^-$, but also for sentences of the form $T(t)$, where t is a term. Hence, such an axiom could look as follows

$$\forall t\ T(\ulcorner T(t)\urcorner) \equiv T(t^\circ). \qquad \text{(TRP)}$$

TRP is meant to abbreviate “TRansParency”. However, we cannot have (TRP) and take the axioms of CT$^-$ with the quantifiers $\forall\varphi$, $\forall\psi$ ranging over all $\mathcal{L}_T$ sentences, since the resulting theory would prove

$$T(\ulcorner\varphi\urcorner) \equiv \varphi,$$

for every sentence φ with the truth predicate, and hence be inconsistent by Tarski's theorem. In fact, as shown by Halbach (1994; see also Halbach, 2011), TRP together with the negation axiom for the whole language $\mathcal{L}_T$, i.e. the sentence

$$\forall \varphi \in \mathrm{Sent}_{\mathcal{L}_T}\ (T(\neg\varphi) \equiv \neg T(\varphi)). \qquad (\mathrm{NEG}(\mathcal{L}_T))$$

are enough to reconstruct the reasoning in the Liar paradox.[4] Hence, we have two ways to go:

1. Accept TRP and resign from (NEG($\mathcal{L}_T$)).
2. Reject TRP and try with compositional axioms à la CT$^-$.

Two truth theories which we shall discuss represent the above two possibilities. The first one is KF, which stands for KripkeFeferman and generalises PT$^-$ to the untyped realm:[5]

Definition 11. KF$^-$ results by adding to PA the following $\mathcal{L}_T$ sentences:

1. $\forall s, t\ \ [T(\ulcorner s = t\urcorner) \equiv s^\circ = t^\circ]$.
2. $\forall s, t\ \ [T(\ulcorner \neg s = t\urcorner) \equiv s^\circ \neq t^\circ]$.
3. $\forall t\ \ [T(\ulcorner T(t)\urcorner) \equiv T(t^\circ)]$.
4. $\forall t\ \ [T(\ulcorner \neg T(t)\urcorner) \equiv [T(\ulcorner \neg t^\circ\urcorner) \vee \neg \mathrm{Sent}_{\mathcal{L}_T}(t^\circ)]$.
5. $\forall \varphi \in \mathcal{L}_T\ [T(\ulcorner \neg\neg\ \varphi\urcorner) \equiv T(\varphi)]$.
6. $\forall \varphi, \psi \in \mathcal{L}_T\ [T(\ulcorner \varphi \wedge \psi\urcorner) \equiv T(\varphi) \wedge T(\psi)]$.
7. $\forall \varphi, \psi \in \mathcal{L}_T\ \ [T(\ulcorner \neg(\ \varphi \wedge \psi)\urcorner) \equiv T(\ulcorner \neg\ \varphi\urcorner) \vee T(\ulcorner \neg\ \psi\urcorner)]$.
8. $\forall v \in \mathrm{Var}\ \forall \varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ [T(\ulcorner \forall v \varphi\urcorner) \equiv \forall t\ T(\varphi[t/v])]$.
9. $\forall v \in \mathrm{Var}\ \forall \varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ [T(\ulcorner \neg\forall v \varphi\urcorner) \equiv \exists t\ T(\ulcorner \neg\varphi[t/v]\urcorner)]$.
10. $\forall t, s\ \forall v \in \mathrm{Var}\ \forall \varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ [s^\circ = t^\circ \rightarrow\ T(\varphi[s/v]) \equiv T(\varphi[t/v])]$.

As usual KF$_n$ results from KF$^-$ by adding induction axioms for Σ_n formulae of $\mathcal{L}_T$ and KF admits induction axioms for all $\mathcal{L}_T$ formulae.

Let us observe that axioms 5.–10. are axioms of PT$^-$ generalised to all formulae and axioms 3.–4. corresponds to axioms 1.–2., giving positive compositional truth conditions for atomic T sentences. Let us mention that this set of axioms is consistent, as essentially shown already by Kripke in (1975). The intuition behind KF$^-$ is that $T(x)$ expresses: "x is a true and grounded sentence", in the

[4] There is a subtlety in here: the contradiction can be derived granted that we have a slightly richer language: symbols representing some primitive recursive functions should be added.

[5] More correctly: PT$^-$ is the typed counterpart of KF$^-$, since the latter theory was formulated first.

sense discussed when describing PT^{-}. Unlike in the typed case, however, this time some sentences are essentially ungrounded, for example the standard liar sentence, φ_L, constructed via the diagonal lemma. One can show that the assumption that it's truth is determined, i.e.

$$T(\ulcorner \varphi_L \urcorner) \vee T(\ulcorner \neg\varphi_L \urcorner)$$

holds, leads to the conclusion that both φ_L and $\neg\varphi_L$ are true, i.e. $T(\ulcorner \varphi_L \urcorner) \wedge T(\ulcorner \neg\varphi_L \urcorner)$ holds. Basing on the intuitive description we gave for PT^{-} it can easily be seen that no grounded sentence can have this property.[6]

Let us now discuss one essential drawback of KF: the theory is essentially asymmetric, in the sense that it's theorems need not be provably true. For example, KF does not prove that the sentence

$$\forall\varphi\ T(\varphi) \vee \neg T(\varphi),$$

which is an easy consequence of KF, is true. Indeed, by using the compositional axioms of KF one can show that

$$T\,(\ulcorner \forall\varphi\ T(\varphi) \vee \neg T(\varphi) \urcorner)$$

is equivalent to

$$\forall\varphi\ T(\varphi) \vee T(\ulcorner \neg\varphi \urcorner),$$

i.e. the assertion that each sentence is determined. What is worse, this theory cannot even be consistently closed under the following rules of reasoning

$$\frac{\varphi}{T(\varphi)}\ \text{(NEC)} \qquad\qquad \frac{T(\varphi)}{\varphi}\ \text{(CONEC)}$$

since one can show that (TRP) (axioms 3. and 4.), axiom 7. and the above two rules are enough for reconstructing the reasoning from the Liar paradox. This essentially shows that if one demands a little bit compositionality and closure under (NEC) and (CONEC) rules, then one should reject one of the axioms 3. or 4. Recall also that TRP is inconsistent with (NEG($\mathcal{L}_T$)). It turns out, that rejecting

[6] One of Kripke's (1975) achievements is to give a formal framework for discussing such issues with mathematical rigour. For example one can distinguish paradoxicality from ungroundedness and conclude e.g. that the truth-teller sentence φ_T, i.e. the sentence asserting its own truth, is ungrounded but not paradoxical. Intuitively this means that the step-by-step procedure described when explaining groundedness does not lead to the ascription of either truth or falsity to φ_T, but the stipulation that its truth is determined does not force us to conclude that it is both true and false.

TRP we can have both closure under (NEC) and (CONEC) and the global axiom for the negation. A theory that goes in this direction is called FS^-:

Definition 12. FS^- results by adding to PA the following axioms:

1. $\forall s, t\ \ [T(\ulcorner s = t\urcorner) \equiv s^\circ = t^\circ]$.
2. $\forall \varphi \in \mathcal{L}_T\ [T(\ulcorner\neg\, \varphi\urcorner) \equiv \neg T(\varphi)]$.
3. $\forall \varphi, \psi \in \mathcal{L}_T\ [T(\ulcorner\varphi\wedge\psi\urcorner) \equiv T(\varphi)\wedge T(\psi)]$.
4. $\forall v \in \mathrm{Var}\ \forall \varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ [T(\ulcorner\forall v\varphi\urcorner) \equiv \forall t\ T(\varphi[t/v])]$.
5. $\forall t, s\ \forall v \in \mathrm{Var}\ \forall \varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ [s^\circ = t^\circ \longrightarrow\ T(\varphi[s/v]) \equiv T(\varphi[t/v])]$.

and closing the resulting theory under the rules (NEC) and (CONEC). The meaning of FS_n and FS are as in the rest of cases, except that the extended theory is also closed under (NEC) and (CONEC).

The above list of axioms is due to Halbach (1994), who offered it as a compositional axiomatization of Friedmann-Sheard theory from (1987). In effect the results of both papers taken together provide a consistency proof for this theory. Let us note that the above axioms are straightforward generalisations of CT^- axioms to the untyped case.

At first sight FS^- seems to be a much more natural theory than KF: it is fully compositional and the lack of (TRP) axioms is somehow compensated by the rules (NEC) and (CONEC). However, there is a high price to be paid: as famously shown by McGee in (1985; see also Halbach, 2011) we cannot find the interpretation for FS^- in the standard model of arithmetic, $\mathbb{N}$. Let us contrast this with the case of e.g. CT^-. While for the latter theory it is true, that no definable subset of $\mathbb{N}$ can serve as its interpretation, there is a set $Tr \subseteq \mathbb{N}$ such that

$$(\mathbb{N}, Tr) \models \mathrm{CT}^-.$$

For Tr it is sufficient to take the set of Gödel codes of all sentences which are true according to the standard Tarskian definition. Similarly, one can find the interpretation for the KF truth predicate in $\mathbb{N}$—this was the way Kripke has originally shown the consistency of KF.[7] No such subset exists for FS^-, as witnessed by the theorem below:

Theorem 13 (McGee, 1985). FS^- *is* ω*−inconsistent, i.e. there is a formula* $\varphi(x) \in \mathcal{L}_T$ *such that* $\mathrm{FS}^- \vdash \exists x\varphi(x)$*, but for every* $n \in \mathbb{N}$, $\mathrm{FS}^- \vdash \neg\varphi(\underline{n})$.

[7] This is an anachronism, since KF was not formulated at that time. However, Kripke's fixpoint construction gives immediately the consistency of this theory.

2.4. Untyped Disquotational Theories

As the Reader probably observed with each new section, we introduce types of theories which are more expressive that the previous ones. Hence presenting disquotational theories at the end may seem to be surprising. However, as the next theorem witnesses this is fully justified. Let PAT denote a theory in $\mathcal{L}_T$ extending PA with no non-logical axioms for the predicate T.

Theorem 14 (McGee, 1992). *For every $\varphi \in \mathcal{L}_T$ there exists a ψ such that* TB(ψ) *is equivalent to φ over* PAT, *i.e.*

$$\mathrm{PAT} \vdash \varphi \equiv (T(\psi) \equiv \psi).$$

The proof of the above is very simple and it is called "McGee's trick": we fix φ and consider the formula

$$\theta(x) \coloneqq \varphi \equiv T(x).$$

By the diagonal lemma there exists a sentence ψ such that

$$\mathrm{PAT} \vdash \psi \equiv \theta(\ulcorner\psi\urcorner).$$

The thesis follows by the definition of θ and the associativity of $\equiv$. As a corollary, if φ is any of the axioms for KF^- or FS^-, then there exists an untyped T-biconditional which is equivalent to it. In other words, both FS^- and KF^- can be presented as untyped disquotational theories.[8]

It can be argued that theories obtained by McGee's trick are not natural and hence shouldn't be taken into consideration, however it is hard to make this argument precise. Now we shall deal with some more natural (in the intuitive sense) theories, which directly correspond to TB and UTB. To define them assume that our language $\mathcal{L}_T$ does not contain the symbol $\rightarrow$ for material implication. If φ is a formula in such a language then we call φ positive if each occurrence of T is within an even number of negations. The assumption that $\rightarrow$ does not occur in φ is important, since $\theta \rightarrow \theta'$ is equivalent to $\neg\theta \vee \theta'$, hence $\rightarrow$ hides one occurrence of the negation.

Definition 15. PTB^- extends PA with axioms TB(ψ) for positive $\mathcal{L}_T$ sentences ψ. PUTB^- extends PA with all sentences of the form UTB(ψ) for positive $\mathcal{L}_T$ formu-

[8] We owe the Reader some explanation, since FS^- was defined as the closure of a certain group of axioms under two rules of reasoning, so the last claim does not exactly follows from the McGee's trick. However, as shown by Halbach (2011), FS^- can be equivalently axiomatized by an infinite set of reflection principles and without the use of additional rules of reasoning.

lae ψ. PTB and PUTB denote the extensions of PTB$^-$ and PUTB$^-$ with induction axioms for all formulae of $\mathcal{L}_T$.

Since every arithmetical formula is trivially positive, PTB$^-$ and PUTB$^-$ extends TB− and UTB− respectively. Moreover, as in the typed case PTB$^-$ is a subtheory of PUTB$^-$ and the latter theory is provable in KF$^-$ (as shown by Cantini, see Halbach, 2011). In particular both theories are consistent. We shall deal with their expressiveness in the last section.

2.5. Additional Axioms

We shall briefly consider some axioms that can be additionally added to the above theories. Each axiom has both typed and untyped version. Let us start with axioms for positive compositional theories (PT$^-$, KF$^-$). Since such theories does not prove axiom (NEG($\mathcal{L}$)), or its untyped variant (NEG($\mathcal{L}_T$)), this gives rise to two principles that can be studied separately. For further usage let us introduce the following abbreviations:

$$\mathrm{tot}(\varphi(v)) \coloneqq \forall t\, (T(\varphi[t/v]) \vee T(\neg\varphi[t/v]))$$
$$\mathrm{cons}(\varphi(v)) \coloneqq \forall t\, \neg(T(\varphi[t/v]) \wedge T(\ulcorner\neg\varphi\, [t/v]\urcorner))$$

and let us define first two principles:

$$\mathrm{COMP} \coloneqq \forall v \in \mathrm{Var}\, \forall\varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ \mathrm{tot}(\varphi)$$
$$\mathrm{CONS} \coloneqq \forall v \in \mathrm{Var}\, \forall\varphi \in \mathrm{Form}^{v}_{\mathcal{L}_T}\ \mathrm{cons}(\varphi).$$

The intended reading of the above principles is that tot (abbreviates “total” and) means that the truth value of $\varphi(t)$ is determined for every term t. cons abbreviates “consistent” and expresses that for no term t, $\varphi(t)$ is contradictory (i.e. both true and false). COMP and CONS abbreviate “complete” and “consistent” respectively. Let us observe that both CONS and COMP are provable in FS$^-$. However neither of them is provable in KF$^-$ (see Halbach, 2011).

The next axioms give us a little bit of induction: each of them will be provable in any of the theory considered in the presence of induction for bounded $\mathcal{L}_T$ formulae. Let us start with the most obvious one, called internal induction. We first define INT(φ) to be the sentence

$$\big(T(\varphi[0/v]) \wedge \forall x(T(\varphi[\underline{x}/v]) \to T(\varphi[\underline{x+1}/v]))\big) \to \forall x\, T(\varphi[\underline{x}/v])$$

$\underline{x}$ denotes the canonical numeral for x (the procedure of assigning canonical numerals to numbers formalizes in PA) and consequently $T(\varphi[\underline{x}/v])$ should be read: “The result of substituting the canonical numeral naming x for every free occurrence of variable v in φ, is true.”

Now we define

$$\forall v \in \mathrm{Var}\ \forall \varphi \in \mathrm{Form}^{v}\ \mathrm{INT}(\varphi). \qquad (\mathrm{INT}(\mathcal{L}))$$

In short, (INT($\mathcal{L}$)) says that every arithmetical formula satisfies induction axiom. In the context of PT^{-} (when it is not provable that every arithmetical formula is total), we shall study also its restriction to total formulae:

$$\forall v\ \forall \varphi \in \mathrm{Form}^{v}\ (\mathrm{tot}(\varphi) \rightarrow \mathrm{INT}(\varphi)). \qquad (\mathrm{INT}_{\mathrm{tot}}(\mathcal{L}))$$

In the context of untyped theories it makes sense to consider these axioms for entire $\mathcal{L}_T$ language. They are denoted INT($\mathcal{L}_T$) and INT$_{\mathrm{tot}}$($\mathcal{L}_T$) and differ from the above axioms only by changing Form^{v} to Form$^{v}_{\mathcal{L}_T}$.

If P is any theory and φ is any additional axiom, then we use the notation

$$\mathsf{P} + \varphi$$

to denote the theory $\mathsf{P} \cup \{\varphi\}$, i.e. the theory resulting from P by adding the axiom φ. In the context of extensions of FS^{-} we also write FS^{-} + φ + (NEC) if we want to consider the closure of FS^{-} + φ under (NEC) rule (+ (CONEC) has an analogous meaning).

* * *

The above short presentation was meant to convince the Reader that there are many ways of axiomatically defining a meaningful notion of truth. We believe that each of the above presented theories, taking into considerations also their variations by adding induction axioms or what we called “additional axioms”, is interesting on its own. In the next section we shall also treat them as examples illustrating differences between various notions of strength.

3. Relations of Strength

In this section we shall study some possibilities of defining order on theories of truth. The intuitive idea will be that a stronger (or more properly: not weaker than) theory “says more” (“no less”) about the property of being true or about the extrasemantical realm of the base theory. Each of the defined relations will be reflexive and transitive (i.e. a preorder), hence will in natural way give rise to an equivalence relation on theories via the condition

$$\mathsf{P} \equiv \mathsf{P'} \iff \mathsf{P} \leq \mathsf{P'} \wedge \mathsf{P'} \leq \mathsf{P},$$

where $\leq$ is the chosen relation. Hence, each of the relations below can also be seen as a way of explicating the notion of "sameness" between theories. Similarly, each such relation$\Longleftrightarrow$ naturally gives rise to a strict preorder relation $\lneq$ via the condition $\mathsf{P} \lneq \mathsf{P}$' $\mathsf{P} \leq \mathsf{P}' \wedge \mathsf{P}' \nleq \mathsf{P}$.

Let us observe that there is one obvious first relation to consider: we say that a theory P is no weaker than P' iff $\mathsf{P} \vdash \mathsf{P}$', i.e. P proves every axiom of P'. Then if P and P' are comparable both ways, i.e. $\mathsf{P} \vdash \mathsf{P}$' and $\mathsf{P}' \vdash \mathsf{P}$ holds, then P and P' are simply axiomatizations of the same set of sentences. This relation provides us with an obvious necessary condition for any other candidate for the relation of strength, for if $\mathsf{P} \vdash \mathsf{P}$' then P should be no weaker than P' in any reasonable sense of "no weaker than". However, for various reasons this relation is too strict being too sensitive to slight perturbations of our theories. We shall now proceed to more permissive candidates and discuss what is here meant by "slight perturbations".

3.1. Fujimoto Definability

This relation, under the name of relative truth definability, was introduced by Kentaro Fujimoto in (2010). To present the way it improves on $\vdash$ let us consider the following slight perturbation of TB^-:

$$\mathsf{P} \coloneqq \mathrm{TB}^- + \forall t\ T(\ulcorner t = t \urcorner),$$

i.e. the extension of TB^- by a single axiom mentioned on the right. As argued when discussing TB^-,

$$\mathrm{TB}^- \nvdash \mathsf{P}.$$

However, TB^- can "improve itself" to make P provable. Consider a formula

$$T'(x) \coloneqq T(x) \vee \exists t\ \ x = \ulcorner t = t \urcorner,$$

which intuitively says that x is either true, in the sense of the truth predicate T, or is a sentence of the form $t = t$, for some closed term t. Then, if φ is any sentence of $\mathcal{L}$, then

$$\mathrm{TB}^- \vdash T'(\ulcorner \varphi \urcorner) \equiv \varphi.$$

Indeed, the right-to-left implication is immediate, since by the very logical form of T' we have (provably in TB^-)

$$\forall x\ T(x) \rightarrow T'(x).$$

The left-to-right one is also easy, since sentences of the form $t = t$ are true. Moreover, by definition,

$$\mathrm{TB}^{-} \vdash \forall t\, T'(\ulcorner t = t \urcorner),$$

hence T', provably in TB^{-} serves as an interpretation for a truth predicate satisfying P. This was one the simplest cases of Fujimoto definability. Now the definition proper:

Definition 16. If Θ, $T'(x)$ are any $\mathcal{L}_T$ formulae, then $\Theta[T'/T]$ denotes a formula in which any occurrence of $T(t)$ is replaced with $T'(t)$ (possibly with renaming bounded variables so as to avoid clashes). If P is any theory of truth, then

$$\mathsf{P}\,[T'/T]$$

denotes the theory whose axioms are sentences $\Theta[T'/T]$ for Θ—an axiom of P. We say that P Fujimoto defines P' if there exists an $\mathcal{L}_T$ formula T' such that

$$\mathsf{P} \vdash \mathsf{P}'\,[T'/T].$$

In such a case we also say that P is Fujimoto no weaker than P' and write $\mathsf{P}' \leq_F \mathsf{P}$. If this holds both ways, then we say that P and P' are Fujimoto equivalent and denote it $\equiv_F$.

Let us observe that if $\mathsf{P} \vdash \mathsf{P}'$ then obviously $\mathsf{P}' \leq_F \mathsf{P}$, for one can take $T'(x) \coloneqq T(x)$ to witness Fujimoto definability. But the example preceding the definition shows that the relation is more permissive and still very intuitive: if $\mathsf{P}' \leq_F \mathsf{P}$, then P admits conceptual resources to prove all about the notion of truth that P' can prove. The fact that it uses a definable predicate T' and does not simply prove that T satisfies axioms of P' should not be that important (for the extended philosophical discussion consult the original Fujimoto's paper [2010]).

Let us point one canonical, easy, but somewhat surprising fact about Fujimoto definability. Consider two theories extending KF^{-} with additional axioms introduced at the end of the last section: KF^{-} + COMP and KF^{-} + CONS. The first one extends KF^{-} with the sentence saying, that the truth value of every sentence is determined, the second one—with the sentence saying that no sentence is both true and false. Observe that since $\mathrm{KF}^{-} + \mathrm{COMP} + \mathrm{CONS} \vdash (\mathrm{NEG}(\mathcal{L}_T))$ then by Tarski's theorem this theory is inconsistent. Hence,

$$\mathrm{KF}^{-} + \mathrm{COMP} \vdash \neg\mathrm{CONS}$$
$$\mathrm{KF}^{-} + \mathrm{CONS} \vdash \neg\mathrm{COMP}.$$

However, consider the formula

$$T'(x) := \neg T(\ulcorner \neg x \urcorner),$$

saying: “$\neg x$ is not true”. It can be easily shown that

$$\mathrm{KF}^{-} \vdash \mathrm{KF}^{-}[T'/T].$$

Moreover, reasoning in KF^{-} the following are equivalent for every sentence $\varphi \in \mathcal{L}_T$

1. $\neg(T'(\varphi) \wedge T'(\ulcorner \neg\varphi \urcorner))$
2. $\neg T'(\varphi) \vee \neg T'(\ulcorner \neg\varphi \urcorner)$
3. $\neg\neg T(\ulcorner \neg\varphi \urcorner) \vee \neg\neg T(\ulcorner \neg\neg\varphi \urcorner)$
4. $T(\ulcorner \neg\varphi \urcorner) \vee T(\varphi)$

the last sentence is an instantiation of COMP for sentence φ. So also 1. is true for φ and since φ was arbitrary we proved CONS$[T'/T]$ in KF^{-} + COMP. The same T' works in the other direction and thus it shows that KF^{-}+ COMP $\leq_F$ KF^{-} + CONS and KF^{-}+ CONS $\leq_F$ KF^{-} + COMP. Hence

$$\mathrm{KF}^{-} + \mathrm{COMP} \equiv_F \mathrm{KF}^{-} + \mathrm{CONS}.$$

Let us now note two important structural properties of Fujimoto definability. Let $\mathrm{Ind}(\mathcal{L}_T) := \{\mathrm{Ind}(\varphi) \mid \varphi \in T\}$ denote the set of induction axioms for $\mathcal{L}_T$ formulae. Then we have

Proposition 17. *If* $\mathsf{P} \leq_F \mathsf{P}'$, *then* $\mathsf{P} + \mathrm{Ind}(\mathcal{L}_T) \leq_F \mathsf{P}' + \mathrm{Ind}(\mathcal{L}_T)$.

The proof of this is almost obvious: if $T'(x)$ witnesses Fujimoto definability of P in P', then the very same formula witnesses Fujimoto definability of P + $\mathrm{Ind}(\mathcal{L}_T)$ in $\mathsf{P}' + \mathrm{Ind}(\mathcal{L}_T)$. This is because for every $\mathcal{L}_T$ formula φ, $\varphi[T'/T]$ is again an $\mathcal{L}_T$-formula, so we have an induction axiom for it in $\mathsf{P}' + \mathrm{Ind}(\mathcal{L}_T)$. As a corollary we get that

$$\mathrm{KF} + \mathrm{CONS} \equiv_F \mathrm{KF} + \mathrm{COMP}.$$

The second property is connected to models. Let us introduce a piece of notation: if $\mathcal{M}$ is any model of some language $\mathcal{L}'$ and $\varphi(x)$ is any $\mathcal{L}'$-formula with at most one free variable, then $\varphi^{\mathcal{M}}$ denotes the set of elements satisfying φ in $\mathcal{M}$. Now suppose $\mathcal{M}$ is a model of PA with universe $\mathcal{M}$ and $Tr \subseteq \mathcal{M}$ is such that $(\mathcal{M}, Tr) \vDash \mathsf{P}'$. If $\mathsf{P} \leq_F \mathsf{P}'$, then

$$(\mathcal{M}, \varphi^{(\mathcal{M},Tr)})$$

is a model of P. Indeed, by Fujimoto definability, we know that $(\mathcal{M}, Tr) \vDash \mathsf{P}[\varphi/T]$ and for every $\psi \in \mathcal{L}_T$ we have

$$(\mathcal{M}, Tr) \vDash \psi[\varphi/T] \iff (\mathcal{M}, \varphi^{(\mathcal{M},Tr)}) \vDash \psi.$$

In particular, if $\mathsf{P} \leq_F \mathsf{P}$', then in any model of P' we can find an interpretation for the truth predicate which makes $\mathcal{M}$ a model of P. By McGee's theorem (Theorem 13) it now follows that FS^- is not Fujimoto definable in any of the theories discussed previously (except FS^- itself, of course), since they all admit an interpretation in the standard model of arithmetic, $\mathbb{N}$.

3.2. Model-Theoretic Strength

With the discussion about models of truth theories from the last section we implicitly moved to the second notion of strength. Let us offer now some more explanation to better prepare for the upcoming definitions.

Models of axiomatic theories of truth are of the form $(\mathcal{M}, Tr)$, where $\mathcal{M}$ is a model of the base theory, which is PA in our case, and Tr is a subset of the universe of $\mathcal{M}$ (denoted M). Most of truth theories (extensions of FS^- being the unique exceptions out of the theories discussed above) can be interpreted in the standard model of arithmetic, $\mathbb{N}$, i.e. we can find a set of natural numbers (codes of some sentences) $Tr \subseteq \mathbb{N}$ such that $(\mathbb{N}, Tr)$ serves as a model of the chosen theory of truth. For typed theories (TB, UTB, PT^-, CT^- and various their extensions) this is unproblematic, for every element of the universe of $\mathbb{N}$ (i.e. a natural number) n such that

$$\mathbb{N} \vDash \text{“}n \text{ is a sentence”}$$

is a code of some real sentence φ. Hence if we put

$$Tr := \{\ulcorner\varphi\urcorner \in \mathbb{N} \mid \mathbb{N} \vDash \varphi\},$$

then $(\mathbb{N}, Tr)$ will be a model of CT, hence any other typed theory of truth discussed in Section 2. Let us observe that in such a model all induction axioms are automatically true, since every nonempty set of natural numbers has the least element ($\mathbb{N}$ is well-founded).

Neither of the above discussed features that helped us defining an extension for CT in $\mathbb{N}$ transfers to non-standard models of PA. These are, by definition, models of PA that are not isomorphic with $\mathbb{N}$, or, equivalently, which contain an element c which is greater than any of their elements denoted by a standard numeral. More precisely: $\mathcal{M}$ is non-standard if it contains an element c such that for every natural number n

$$\mathcal{M} \vDash c \neq \underline{n}.$$

(Recall that $\underline{n}$ denotes the canonical numeral naming n.) Every such element c is called non-standard. It immediately follows that, in any non-standard model of PA, there must be infinitely many such elements, since neither of

$$c + 1, c + 2,\ldots, c - 1, c - 2,\ldots, 2c, 3c,\ldots, c^2, c^3,\ldots$$

can be equal to a standard (i.e. not non-standard) number. In particular non-standard models of PA are never well-founded, since if c is any non-standard element, then so are

$$c - 1 > c - 2 > c - 3 > \ldots > c - n > \ldots$$

for every n. The existence of such models follows from two independent sources: the fact that PA is incomplete (by Gödel's first incompleteness theorem) and the compactness theorem for first-order logic. Kaye (1991) provides a very good first introduction to non-standard models of PA.

One might be tempted to ignore the existence of such bizarre objects as described above. After all, don't we know that natural numbers are well-founded and hence we can limit our attention to what happens in ℕ? However, this criterion is stated in a metatheory with respect to PA and cannot be restated within this system: for PA all its models are on a par or at least there seems to be no convincing internal to PA arguments ruling out ill-founded models.[9] This point of view (that all models are on a par) is particularly adequate if our base theory is a toy model of all of our knowledge about extrasemantical facts (e.g. those that do not involve the notion of truth). Such a perspective is often adopted e.g. in the formal study of the deflationism (see e.g. Fischer & Horsten, 2015). Hence, if the base theory is all that we know, all its models should be equally good for us. We think about them as of possible worlds which are admissible from the point of view of our theory. The main question now is: when such a possible world can be expanded to a model of the respective truth theory?

Definition 18. Let $\mathcal{M} \models \mathrm{PA}$. We say that $\mathcal{M}$ expands to a model of an $\mathcal{L}_T$ theory P if there exists $Tr \subseteq M$ such that $(\mathcal{M}, Tr) \models \mathsf{P}$.

Let us now show why, for a given non-standard model $\mathcal{M} \models \mathrm{PA}$, the question whether $\mathcal{M}$ expands to a model of a theory of truth P is, in case of many P's, non-trivial. Why can't we simply do what we did in the case of ℕ and take $Tr_{\mathcal{M}} := \{\ulcorner\varphi\urcorner \ | \ \mathcal{M} \models \varphi\}$? This, in fact, will work but only for one of our theories: TB$^-$. It is a very simple exercise to show, that for every $\mathcal{M}$

[9] In (2018) Cieśliński argued that implicit commitments of theories may provide us with an internal to PA (or any base theory) ways of separating some non-standard models. However, even in this approach, all non-standard models that we will need later on, are left untouched.

$$(\mathcal{M}, Tr_{\mathcal{M}}) \vDash \mathrm{TB}^{-}.$$

This however breaks down already for UTB^{-}. As we have already noted, PA proves that for every x there exists a canonical numeral naming x. This numeral, provably in PA consists of a string of S's of length x followed by 0 (with some brackets in the meantime, compare (Num)). Since this is a theorem of PA, then if $\mathcal{M}$ is any non-standard model with a non-standard element c, then

$$\mathcal{M} \vDash \text{“There exists a canonical numeral naming } c\text{”}$$

Let us fix the unique element $\underline{c}$ witnessing the above statement. Then

$$\mathcal{M} \vDash \text{“}\underline{c} \text{ consists of a string of } S\text{'s of length } c \text{ followed by 0”}.$$

In particular $\underline{c}$ itself is a non-standard element of $\mathcal{M}$. Let us now suppose that for some set $A \subseteq M$, $(\mathcal{M}, A) \vDash$ UTB^{-}. Since $\mathcal{M} \vDash c = c$, and $\underline{c}$ is a term naming c, it follows that $\mathcal{M} \vDash \underline{c}^{\circ} = \underline{c}^{\circ}$. Hence by our assumption we must have that $\ulcorner \underline{c} = \underline{c} \urcorner$ (i.e. the code of a sentence asserting that $c = c$) is an element of A. However $\ulcorner \underline{c} = \underline{c} \urcorner \notin Tr_{\mathcal{M}}$ because it is a non-standard element (since already $\underline{c}$ is).[10]

This, however, can easily be fixed: for a given nonstandard model $\mathcal{M}$, define

$$UTr_{\mathcal{M}} := \{\ulcorner \varphi(t_1,\dots,t_n)\urcorner \in M \ \mid \ \varphi(x_1,\dots,x_n) \in \mathrm{Form}_{\mathcal{L}}, \mathcal{M} \vDash \varphi(a_1,\dots,a_n) \wedge \bigwedge_{i \leq n} a_i = t_i^{\circ}\}.$$

$UTr_{\mathcal{M}}$ consists of codes of those formulae which are of standard logical depth, but may contain arbitrary terms, and are true in the model on the values of these terms (according to the model). This set, however, stops working even for the simpler out of our two typed compositional theories: PT^{-}. To see this, observe e.g. that PA proves that for every x there exists a sentence consisting of a string of x negations and then $0 = 0$ i.e. the sentence of the form

$$\underbrace{\neg\dots\neg}_{x \text{ times } \neg} 0 = 0$$

Hence in any non-standard model $\mathcal{M}$ and for any non-standard element $c \in M$ we must have a sentence $\theta \in M$ such that

$$\mathcal{M} \vDash \text{“}\theta \text{ starts with a string of } c \text{ negations which is followed by } 0 = 0\text{”}.$$

[10] We tacitly assume that our coding is monotone, i.e. a sequence has always no smaller code than any its subsequence. Most standard Gödel codings have this property.

Now suppose $(\mathcal{M}, A) \vDash \mathrm{PT}^{-}$ and θ and c are as in the above. Then, by axiom 1. of PT^{-}, $\ulcorner \neg 0 = 1 \urcorner \in A$, since $\mathcal{M} \vDash 0 \neq 1$. Now let us consider a sentence (in the sense of $\mathcal{M}$) $\ulcorner \neg(\theta \wedge 0 = 1) \urcorner \in M$. By axiom 5., which expresses that the negation of conjunction is true iff one of the conjuncts is false, we have

$$(\mathcal{M}, A) \vDash T(\ulcorner \neg(\theta \wedge 0 = 1)\urcorner) \equiv T(\ulcorner \neg \theta \urcorner) \vee T(\ulcorner \neg 0 = 1 \urcorner).$$

Hence, since $\ulcorner \neg 0 = 1 \urcorner \in A$, also $\ulcorner \neg(\theta \wedge 0 = 1)\urcorner \in A$. But $\ulcorner \neg(\theta \wedge 0 = 1)\urcorner \notin UTr_{\mathcal{M}}$, since this sentence does not have a standard logical depth (already θ contains a non-standard number of negations). Hence, in a non-standard model $\mathcal{M}$, $UTr_{\mathcal{M}}$ is never an interpretation for PT^{-}.

For PT^{-} this too can be fixed, but requires a slightly more complicated argument (one formalizes the step-by-step procedure described in the previous section and iterates it through ordinal numbers). Hence, for every non-standard model $\mathcal{M}$ there exists a set $PTr_{\mathcal{M}}$ such that $(\mathcal{M}, PTr_{\mathcal{M}}) \vDash \mathrm{PT}^{-}$. What about the next step and the case of CT^{-}? Here, in order to find an extension for it in a non-standard model $\mathcal{M}$, we encounter another level of complication: for any two sentences φ, $\neg\varphi$ in the sense of $\mathcal{M}$ we have to choose exactly one and put it into the extension. It can be shown, that the procedure of finding $PTr_{\mathcal{M}}$ leaves the truth value of many non-standard sentences undetermined. For example, if θ is as above, neither θ, nor $\neg\theta$ will be in $PTr_{\mathcal{M}}$. This is essentially due to the well-foundedness of the construction of $PTr_{\mathcal{M}}$.

It turns out that, in general, this cannot be fixed: there are non-standard models of PA in which there is no appropriate extension for CT^{-}. Moreover this is not due to the fact that CT^{-} proves an arithmetical sentence which is not provable in PA, because it does not. One can show that any consistent extension of PA has both a model in which one can interpret CT^{-} and a model in which one cannot do it. We will say more about it when discussing concrete results in Section 4.

We have just seen that the obstacles for finding the interpretation of a theory of truth in a non-standard model may be caused by typically truth-theoretical axioms (such as (NEG($\mathcal{L}$))). Let us mention one more such source of complications: adding $\mathrm{Ind}(\mathcal{L}_T)$ to our theory, i.e. extending the truth theory with all induction axioms for $\mathcal{L}_T$ formulae. Let us illustrate this with the simplest possible case: we shall show that $(\mathcal{M}, UTr_{\mathcal{M}})$ is never a model of TB (mind the lack of “ $^{-}$ ”). Assume the contrary. Let $\varphi(x)$ be the $\mathcal{L}_T$ formula

“Every sentence of logical depth at most x is either true or false”.

Then, by the definition of $UTr_{\mathcal{M}}$, it follows that $(\mathcal{M}, UTr_{\mathcal{M}}) \vDash \varphi(\underline{n})$ for every natural number n. Moreover this reverses: $(\mathcal{M}, UTr_{\mathcal{M}}) \vDash \varphi(a)$ if and only if a is a standard element of $\mathcal{M}$, because there are no elements of non-standard logical depth in $UTr_{\mathcal{M}}$. Consequently,

$$(\mathcal{M}, UTr_{\mathcal{M}}) \vDash \neg\varphi(a) \Longleftrightarrow a \text{ is a non-standard element of } \mathcal{M}.$$

But as we have already seen, the set of all non-standard elements in any non-standard model of PA does not have the least element. Hence the induction axiom for $\neg\varphi(x)$ is not true in $(\mathcal{M}, UTr_{\mathcal{M}})$ and we obtained the desired contradiction. It transpires that, as in the case of CT^{-}, in general this cannot be fixed by improving $UTr_{\mathcal{M}}$. In some non-standard models of PA one simply cannot find the interpretation for the truth predicate satisfying TB. *A fortiori*, this is true about all the rest of truth theories extended with induction axioms for $\mathcal{L}_T$.

We shall base the next relation of strength on the above considerations. Intuitively, a theory is stronger if it puts more restrictions on the universe of a model of our base theory. This should be reflected in the fact that, if P is properly stronger than P', then less models of PA expands to a model of P, then to a model of P'. The definition follows:

Definition 19. We say that P is model-theoretically not weaker than P', and denote it P' $\leq_M$ P, if every model of PA that can be expanded to a model of P, can be expanded to a model of P' as well.

Hence a theory P is model theoretically stronger than P', denoted P' $\lneq_M$ P if P' $\leq_M$ P and P $\nleq_M$ P'. In this situation there are strictly more models of PA that carry an interpretation for P' than those that carry an interpretation for P (i.e. the class of models expandable to P is a proper subclass of the class of models expandable to P').

Let us observe that Fujimoto definability is a refinement of the above relation, in the sense that for all truth theories P and P' it holds that

$$\mathsf{P} \leq_F \mathsf{P'} \Rightarrow \mathsf{P} \leq_M \mathsf{P'}.$$

Indeed, the crucial fact was already observed at the end of the section devoted to Fujimoto definability: if P $\leq_F$ P', then there is a formula (witnessing Fujimoto definability) which in any model of P' defines an interpretation for P. To visualize the difference between the two relations of strength let us write the above condition formally:

$$\exists\varphi(x) \in \mathcal{L}_T\ \forall(\mathcal{M}, A)\ \ ((\mathcal{M}, A) \vDash \mathsf{P'} \rightarrow (\mathcal{M}, \varphi^{(\mathcal{M},A)}) \vDash \mathsf{P}).$$

One can now think about this $\varphi(x)$ as of a fixed procedure for transforming models of P' into models of P. On our way leading to the second notion let us weaken this condition by allowing this procedure to depend on the model $(\mathcal{M}, A)$:

$$\forall(\mathcal{M}, A)\ \ ((\mathcal{M}, A) \vDash \mathsf{P'} \rightarrow \exists\varphi(x) \in \mathcal{L}_T\ \ (\mathcal{M}, \varphi^{(\mathcal{M},A)}) \vDash \mathsf{P}).$$

The above condition permits to pick a different $\varphi(x)$ for every model $(\mathcal{M}, A)$.[11] Now, our second relation relaxes the above condition even further: we do not require that the interpretation of the truth predicate satisfying P be given by a formula. We demand only its existence as a subset of $\mathcal{M}$. Hence formally, the condition for $\mathsf{P} \leq_M \mathsf{P}$' is

$$\forall(\mathcal{M}, A) \quad ((\mathcal{M}, A) \models \mathsf{P}' \rightarrow \exists A' \subseteq M \quad (\mathcal{M}, A') \models \mathsf{P}).$$

It turns out that this shift makes it possible to compare both TB and UTB (which admits full induction) with CT^- (which has no induction for formulae with the truth predicate) and also some untyped truth theories with the typed ones. We shall say more about the concrete results in the last section.

3.3. Proof-Theoretic Strength

The last relation of strength takes us back to the purely syntactical world. It can be best seen as a weakening of the provability relation that we started from: $\mathsf{P} \vdash \mathsf{P}'$ means that

$$\text{For every sentence } \varphi \text{ of } \mathcal{L}_T, \text{ if } \mathsf{P}' \vdash \varphi, \text{ then } \mathsf{P} \vdash \varphi.$$

Now, we want to restrict the above universal condition to $\mathcal{L}$ (i.e. arithmetical) sentences only. Here is our definition:

Definition 20. We say that P is proof-theoretically not weaker than P', denoted $\mathsf{P}' \leq_P \mathsf{P}$, if every arithmetical sentence provable in P' is provable also in P.

Hence “proof-theoretically stronger than” means simply “proves more arithmetical sentences than”. There is a common intuition behind this and the above model-theoretical notion of strength: a stronger theory is one, which says more about the realm of the base theory. Slightly more concretely: a stronger theory is the one which excludes more possibilities admitted by the base theory. In the previous example, these “possibilities” were models of PA and truth theories could exclude some of them by imposing more requirements on their structure. In the current one, the possibilities are sentences which were left undecided by the base theory. Let us observe that the latter condition is no stronger than the former, i.e. for all P, P' it holds that

$$\mathsf{P}' \leq_M \mathsf{P} \Rightarrow \mathsf{P}' \leq_P \mathsf{P}.$$

[11] One can show that the two conditions are equivalent if P is finitely axiomatizable modulo PA, so contains only finitely many specific axioms for the predicate T.

Indeed, let us show the contraposition. Assume that $\mathsf{P}' \not\leq_P \mathsf{P}$, hence there exists a sentence $\varphi \in \mathcal{L}$ such that $\mathsf{P}' \vdash \varphi$ but $\mathsf{P} \nvdash \varphi$. Hence $\mathsf{P} + \neg\varphi$ is a consistent theory. Let $(\mathcal{M}, A) \vDash \mathsf{P} + \neg\varphi$. Since $\neg\varphi$ is arithmetical, $\mathcal{M} \vDash \neg\varphi$. It follows that every expansion of $\mathcal{M}$ satisfies $\neg\varphi$ (because expansions leave the arithmetical part untouched). Hence $\mathcal{M}$ cannot be expanded to a model of P', because the latter theory proves φ. Hence there is a model of P which cannot be expanded to a model of P'. So, $\mathsf{P}' \not\leq_M \mathsf{P}$. That this is in fact a properly weaker condition, will be shown in the next section.

* * *

Let us briefly summarize the developments of this section. We have introduced three relations, denoted $\leq_F$, $\leq_M$, $\leq_P$, such that each of them is a preorder on the set of axiomatic theories of truth. Each of them can be treated as an explication of the notion of "strength" of an axiomatic theory. These relations (as sets of ordered pairs of theories) in turn can be ordered by inclusion, giving us the following picture: for all P, P' it holds that

$$\mathsf{P} \leq_F \mathsf{P}' \Rightarrow \mathsf{P} \leq_M \mathsf{P}' \Rightarrow \mathsf{P} \leq_P \mathsf{P}'. \tag{FMP}$$

Below we shall see that each of these implications is strict. The contraposed picture will be useful later on:

$$\mathsf{P} \not\leq_P \mathsf{P}' \Rightarrow \mathsf{P} \not\leq_M \mathsf{P}' \Rightarrow \mathsf{P} \not\leq_F \mathsf{P}'. \tag{$\neg P \neg M \neg F$}$$

4. Results & Open Problems

In this section we present the known results about the strength of axiomatic truth theories introduced in Section 2. We shall proceed in the reverse order to the one adopted in the previous section, starting from the least restrictive notion of strength. Hence each next subsection can be treated as a refinement of the informations about strength gathered in the previous sections.

4.1. Results on the Proof-Theoretical Strength

Let us observe that the place of each theory P in this hierarchy is determined by the set of its arithmetical consequences, i.e. the set $\mathrm{Pr}_{\mathcal{L}}(\mathsf{P}) := \{\varphi \in \mathcal{L} \mid \mathsf{P} \vdash \varphi\}$. Let us start with saying which theories are the weakest in this sense. Since each of selected theories extends PA, then the least possible choice of $\mathrm{Pr}_{\mathcal{L}}(\mathrm{P})$ is $\mathrm{Pr}_{\mathcal{L}}(\mathrm{PA})$, i.e. the set of arithmetical consequences of PA.We say that such theories are proof-theoretically conservative over PA. Determining which theories satisfy this criterion is one of the main research goals in axiomatic truth theories and the boundary separating the conservative theories from the nonconservative ones is called the Tarski Boundary. (Łełyk, Wcisło, 2017) is devoted mainly to this area.

Theorem 21. *The following theories are proof-theoretically conservative over* PA: UTB, PTB, FS^{-} + INT($\mathcal{L}$) , KF^{-} + CONS + $\mathrm{INT}_{\mathrm{tot}}(\mathcal{L}_T)$.

The conservativity of

1. UTB is due essentially already to Tarski (1995). A proof of it is presented in (Halbach, 2011).
2. PTB is due to Cieśliński (2011).
3. FS^{-} + INT($\mathcal{L}$) is a corollary of the proof of conservativity of CT^{-} + INT($\mathcal{L}$) (originally by Krajewski, Kotlarski and Lachlan, 1981) and Halbach's observation about the relation between FS and $\mathrm{RT}_{<\omega}$ (to be found in Halbach, 2011). (Enayat, Łełyk & Wcisło, 2019) contains an outline of this proof. Note that in this theory we have INT($\mathcal{L}$) and not INT($\mathcal{L}_T$).
4. KF^{-} + CONS + $\mathrm{INT}_{\mathrm{tot}}(\mathcal{L}_T)$ is due to Cantini (1989).

Let us observe that from Theorem 21 it follows that each of

1. TB^{-}, UTB^{-}, TB,
2. PT^{-} (+INT($\mathcal{L}$)), CT^{-} (+INT($\mathcal{L}$)),
3. PUTB^{-}, KF^{-}, KF^{-} + CONS

is proof-theoretically conservative over PA, being a subtheory of some theory listed in Theorem 21 (theories in 1.—of UTB, in 2.—of FS^{-} + INT($\mathcal{L}$), in 3.—of KF^{-} + CONS + $\mathrm{INT}_{\mathrm{tot}}(\mathcal{L}_T)$). Since KF^{-} + COMP is Fujimoto definable in KF^{-} + CONS (as we remarked at the end of Section 3.1) it is also proof-theoretically conservative.

Contemplating the above picture, one can see a common pattern: typed, disquotational theories are conservative (even in the presence of full induction for $\mathcal{L}_T$) and so are pure (i.e. without induction axioms) untyped theories (we note that, however, by McGee's trick, untyped disquotational theories can be arbitrarily strong). A natural question now is: what happens to compositional theories in the presence of induction? Moreover, we haven't commented the status of PUTB, yet. So let us start with the weakest compositional theory from our list, i.e. PT^{-}, and add to it induction for $\mathcal{L}_T$ formulae from the bottom of arithmetical hierarchy (i.e. bounded ones). It turns out that in the presence of even so modest resources, axiom (NEG($\mathcal{L}$)) becomes provable:[12]

[12] Let us observe that the most straightforward argument requires Π_1 induction.

Theorem 22. (Ł., 2017) $\mathrm{PT}_0 \vdash \mathrm{CT}_0$.

Hence, it is not hard to see that actually both theories are equivalent (since obviously $\mathrm{CT}_0 \vdash \mathrm{PT}_0$). It was one of the most important messages of (Łełyk, Wcisło, 2017) that CT_0 is not proof-theoretically conservative over PA. Hence even weak compositionality together with basic induction axioms is enough to make the transition from the conservative to the nonconservative side of the Tarski Boundary. Now one can invoke standard facts known from fragments of arithmetic (see e.g. Hájek & Pudlák, 1993) to conclude that adding induction axioms to CT^- for every next level of arithmetical hierarchy gives us proof-theoretically stronger theories.

Theorem 23 (Folklore). *For every n,* $\mathrm{CT}_n \lneq_P \mathrm{CT}_{n+1}$ *and, consequently,* $\mathrm{CT}_n \lneq_P \mathrm{CT}$.

The characterization of arithmetical consequences of CT_n's and CT is due to Kotlarski and Ratajczyk (1990).

The next step is obviously to consider untyped theories. Here, we also experience a significant rise in strength. Before going further it might be a good exercise to go back to the definition of FS and KF and try to guess which one of them is capable of proving more arithmetical sentences.

Theorem 24 (Friedman-Sheard, 1987; Halbach, 1994). $\mathrm{CT} \lneq_P \mathrm{FS}$.

The next natural question to ask is about subsystems of FS based on restricted induction. Here the answer is particularly simple: for every n, FS_n is equivalent to full FS. Indeed, it suffices to observe that already $\mathrm{FS}_0 \vdash \mathrm{FS}$. This is so, because

$$\mathrm{FS}_0 \vdash \mathrm{INT}(\mathcal{L}_T),$$

and $\mathrm{INT}(\mathcal{L}_T)$ is over FS^- equivalent to the sentence saying "All axioms of induction for $\mathcal{L}_T$ are true". Hence if φ is any instantiation of the axioms of induction for a $\mathcal{L}_T$ formula, then

$$\mathrm{FS}_0 \vdash T(\ulcorner\varphi\urcorner).$$

Then, one application of (CONEC) finishes the argument.[13] In fact we conjecture that this can be improved.[14]

[13] We are grateful to Graham Leigh for pointing this out.

[14] We found the proof of this conjecture while preparing this paper. Since it was not properly verified, we leave it as a conjecture.

Conjecture 25. CT $\lneq_P$ FS$^-$ + INT($\mathcal{L}_T$) + (NEC).

Going back to what is known, let us mention that FS outstrips CT in strength quite significantly: it has the same arithmetical consequences as ω-iterations of CT (i.e. the theory RT$_{<\omega}$. This is a result due to Halbach [1994], see also Halbach, 2011).

We are left with subsystems of KF. It turns out that already a very limited induction axiom is needed to go well beyond full FS:

Theorem 26. FS $\lneq_P$ KF$^-$ + INT($\mathcal{L}_T$).

This is a corollary to a result of Cantini from (1989) and the already mentioned results due to Friedman-Sheard and Halbach. (Nicolai, 2018) gives a very good overview of strength of various KF-like theories. As in the case of CT, the arithmetical strength of fragments of KF can be read of the results on fragments of PA.

Theorem 27 (Folklore). *For every n,* KF$_n$ $\lneq_P$ KF$_{n+1}$ *and, consequently,* KF$_n$ $\lneq_P$ KF.

Moreover, it is known (by the results of Cantini [1989] and Feferman [1991]) that KF + CONS $\equiv_P$ KF. Hence, also KF + COMP $\equiv_P$ KF (by Fujimoto definability).

The last question is: what about PUTB? Given that disquotational theories (UTB, PTB) so far turned out to be weak, one can guess that maybe this is also the case of PUTB. However, let us note that, by McGee'a trick (Theorem 14), for untyped disquotational theories sky is the limit: such theories can be as strong as possible (up to the ultimate limit of proof-theoretical strength: inconsistency). Quite surprisingly, already PUTB takes us very far:

Theorem 28 (Halbach, 2011). KF $\leq_F$ PUTB.

Since KF $\vdash$ PUTB, we also have KF $\equiv_F$ PUTB (note that we are dealing here with Fujimoto definability). In particular, by (FMP), KF $\equiv_P$ PUTB.

Hence we have the following picture summarizing the results on the proof-theoretical strength of theories in consideration:

Conservative theories $\lneq_P$ PT$_0$ = CT$_0$ $\lneq_P$ CT$_1$ $\lneq_P$ … $\lneq_P$ CT $\lneq_P$ FS $\lneq_P$ KF$^-$ + INT($\mathcal{L}_T$) $\lneq_P$ KF$_0$ $\lneq_P$ KF$_1$ $\lneq_P$ … $\lneq_P$ KF $\equiv_P$ PUTB.

4.2. Results on the Model-Theoretical Strength

Let us start by observing that Theorem 21 makes all the various truth theories mentioned in its thesis (together with theories provable in them, mentioned after the theorem) identical with respect to the proof-theoretical strength. It turns out that the model-theoretical notion of strength has the appropriate granularity to distinguish between (almost all of) them. The following diagram summarizes what we know (→ stands for $\lneq_M$, while ⇒ for $\leq_M$):

Figure 1

$$\begin{tikzcd}
 & \mathrm{FS}^- + \mathrm{INT}(\mathcal{L}) & \\
\mathrm{FS}^- \arrow[ur, Rightarrow] & \mathrm{CT}^- + \mathrm{INT}(\mathcal{L}) \arrow[u] & \\
\mathrm{CT}^- \arrow[u] \arrow[ur, Rightarrow] & & \mathrm{PT}^- + \mathrm{INT}(\mathcal{L}) \arrow[ul, Rightarrow] \\
 & \mathrm{UTB} \arrow[ul, Rightarrow] \arrow[ur, Rightarrow] & \\
 & \mathrm{PT}^- + \mathrm{INT}_{\mathrm{tot}}(\mathcal{L}) \arrow[u] & \\
 & \mathrm{PTB} \arrow[u] & \\
 & \mathrm{TB} \arrow[u] & \\
 & \mathrm{PT}^-, \mathrm{TB}^-, \mathrm{UTB}^-, \mathrm{KF}^- + \mathrm{CONS}, \mathrm{KF}^- + \mathrm{COMP} \arrow[u] &
\end{tikzcd}$$

As the first comment, let us say that we do not know whether any of ⇒ is actually →. In particular, we can show that any model which expands to a model of CT$^-$ expands also to a model of UTB (mind the lack of " $^-$ "; this is a result due to Bartosz Wcisło from [Łełyk & Wcisło, 2017]), but we do not know whether this reverses. Similarly we do not know any relations between CT$^-$ and PT$^-$ + INT($\mathcal{L}$).

Let us now explain this diagram in slightly more details. The bottom of this hierarchy is occupied by theories which are model-theoretically as weak as possible: every model of PA expands to a model of PT$^-$ as discussed already in Section 3.2. The same holds for UTB$^-$ and TB$^-$ (they are in fact both provable in PT$^-$) and, by the result of Cantini (see Cantini, 1989; it was implicit already in Kripke's fixpoint construction from [Kripke, 1975]), KF$^-$ with either of COMP and

CONS. The theory at the second level, TB, already eliminates some models. This follows by the result of (independently) Engström and Cieśliński (see Łełyk & Wcisło, 2017 for a proof). Next two arrows are unpublished results by Bartosz Wcisło and the author. That PT$^-$ + INT$_{\text{tot}}(\mathcal{L})$ is strictly between TB and UTB, and UTB is not stronger that PT$^-$ + INT$(\mathcal{L})$ follows from the results of (Łełyk & Wcisło, 2019). The strictness of the topmost (vertical) arrow is the consequence of ω-inconsistency of FS$^-$ + INT$(\mathcal{L})$. Moreover we conjecture that FS$^-$ is model-theoretically incomparable with PT$^-$ + INT$(\mathcal{L})$.

4.3. Results on Fujimoto Definability

Now, we show that most of theories which turned out to be equivalent using two previous notions of strength, can now be distinguished using Fujimoto definability. We shall use the structural properties of Fujimoto definability.

Proposition 29. TB$^- \lneqq_F$ UTB$^- \lneqq_F$ PT$^- \lneqq_F$ KF$^-$

Proof. In all the cases $\leq_F$ follows straightforwardly, since the theory to the left is always provable in its right neighbour. To prove $\ngeq_F$ we look at inductive versions of the above theories and observe that

$$\text{TB} \ngeq_F \text{UTB} \ngeq_F \text{PT} \ngeq_F \text{KF}.$$

The first $\ngeq_F$ follows from the fact that TB $\ngeq_M$ UTB, as noted in Section 4.2. The next two follows from the proof-theoretical strength of respective theories: we have

$$\text{UTB} \ngeq_P \ \text{PT} \ngeq_P \text{KF}.$$

(We used the fact that PT = CT).

There are two more truth theories which are mentioned at the bottom of the diagram from the previous section: KF$^-$ + CONS and KF$^-$ + COMP. As we have already seen they are Fujimoto equivalent. The problem whether each of them is Fujimoto definable in KF alone is open.

To end this section we would like to give some example contrasting Fujimoto definability with the previous two notions. Let us recall that we have

$$\text{PTB} \leq_M \text{PT}^- + \text{INT}(\mathcal{L}), \text{ and UTB} \leq_M \text{CT}^-.$$

In contrast, neither of the inequalities holds for Fujimoto definability (in fact in the first case we have an even stronger negative result):

$$\text{PTB} \nleq_F \text{CT and UTB} \nleq_F \text{CT}^-.$$

Both results are yet unpublished: the second one is due to Albert Visser, while the first one is our observation.

5. Summary

The main aim of our paper was to present three formal tools for comparing various axiomatic theories of truth. In Section 2 we aimed at showing that there are indeed many different approaches to defining a set of axioms for the notion of truth. In Section 3 we introduced three different "measures of strength" of axiomatic theories of truth, i.e. three reflexive and transitive relations (preorders) on the set of axiomatic theories of truth. We have explained the intuition behind each of them. The three relations were called (from the most fine-grained to the coarsest): Fujimoto definability, model-theoretical strength, proof-theoretical strength. Then in the last section we described how they order the truth theories introduced in Section 2. We observed that theories made equivalent by the coarser relation can be strictly ordered by the next one.

REFERENCES

Cantini, A. (1989). Notes on Formal Theories of Truth. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, *35*(1), 97–130.

Cieśliński, C. (2018). The Epistemic Lightness of Truth: Deflationism and its Logic. Cambridge: Cambridge University Press.

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

Enayat, A., Łełyk, M. and Wcisło, B. (2019). Truth and Feasible Reducibility. *The Journal of Symbolic Logic*, 1–58. doi:10.1017/jsl.2019.24.

Feferman, S. (1991). Reflecting on Incompleteness. *The Journal of Symbolic Logic*, *56*(1), 1–49.

Fischer, M. and Horsten, L. (2015). The Expressive Power of Truth. *Review of Symbolic Logic*, *8*(2), 345–369.

Friedman, H. and Sheard, M. (1987). An Axiomatic Approach to Self-Referential Truth. *Annals of Pure and Applied Logic*, *33*, 1–21.

Fujimoto, K. (2010). Relative Truth Definability of Axiomatic Truth Theories. *Bulletin of Symbolic Logic*, *16*(3), 305–344.

Hájek, P. and Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. New York: Springer-Verlag.

Halbach, V. (1994). A System of Complete and Consistent Truth. *Notre Dame J. Formal Logic*, *35*(3), 311–327.

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

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

Kotlarski, H., Krajewski, s. and Lachlan, A. (1981). Construction of Satisfaction Classes for Nonstandard Models. *Canadian Mathematical Bulletin*, *24*, 283–93.

Kotlarski, H. and Ratajczyk, Z. (1990). More on Induction in the Language with a Satisfaction Class. *Mathematical Logic Quarterly*, *36*(5), 441–454.

Kripke, s. (1975). Outline of a Theory of Truth. *The Journal of Philosophy*, *72*(19), 690–716.

Łełyk, M. (2017). Axiomatic Theories of Truth, Bounded Induction and Reflection Principles. PhD thesis, University of Warsaw.

Łełyk, M. and Wcisło, B. (2017). Models of Weak Theories of Truth. *Archive for Mathematical Logic*, *56*(5), 1–26. doi:10.1007/s00153-017-0531-1.

Łełyk, M. and Wcisło, B. (2017). Strong and Weak Truth Principles. *Studia Semiotyczne—English Supplement*, *29*, 107–126.

Łełyk, M. and Wcisło, B. (2019). Models of Positive Truth. *The Review of Symbolic Logic*, *12*(1), 144–172.

McGee, V. (1985). How Truthlike Can a Predicate Be? A Negative Result. *Journal of Philosophical Logic*, *14*(4), 399–410.

McGee, V. (1992). Maximal Consistent Sets of Instances of Tarski's Schema (t). *Journal of Philosophical Logic*, *21*(3), 235–241.

Nicolai, C. (2018). Provably True Sentences Across Axiomatizations of Kripke's Theory of Truth. *Studia Logica*, *106*(1), 101–130.

Tarski, A. (1995). Pojęcie prawdy w językach nauk dedukcyjnych. In: J. Zygmunt (Ed.), *Pisma logiczno-filozoficzne, t. 1: Prawda* (pp. 228–282). Warszawa: PWN.