THE BULLETIN OF SYMBOLIC LOGIC
Volume 16, Number 3, Sept. 2010

RELATIVE TRUTH DEFINABILITY OF AXIOMATIC TRUTH THEORIES

KENTARO FUJIMOTO

Abstract. The present paper suggests relative truth definability as a tool for comparing conceptual aspects of axiomatic theories of truth and gives an overview of recent developments of axiomatic theories of truth in the light of it. We also show several new proof-theoretic results via relative truth definability including a complete answer to the conjecture raised by Feferman in [13].

§1. Introduction. A large variety of axiomatic theories of truth have been proposed. Different theories embody different views and conceptions of truth behind their axioms. Thus it should be expected that one can gain both mathematically and philosophically valuable insights about truth by investigating how they are related to each other and by comparing them from various points of view.

Indeed much proof-theoretic work on axiomatic theories of truth has been done, and many theories have been related to certain subsystems of second-order arithmetic in terms of proof-theoretic equivalence, equiconsistency, etc (see [20] or [37] for an overview). However, in many cases, these works focus only on the arithmetical parts of theories, i.e., their theorems containing no truth predicate.

If one adopts an instrumentalist approach to truth, in which truth is regarded just as practical device to increase the deductive or expressive power of theories, it might be sufficient to consider only what kind of and how rich truth-free consequences can be obtained by means of a truth predicate (or truth predicate*s* if the language has more than one truth predicates). However, if we are interested in more conceptual aspects of truth theories, it seems no longer sufficient to consider only the derivability of truth-free sentences. For, different theories based on different and incompatible views of truth may have the same truth-free consequences just by accident. Hence, in order to compare conceptual aspects of theories of truth, we must take into consideration their theorems which may contain truth predicate(s). We

Received February 13, 2009.

1079-8986/10/1603-0001/$5.00

need mathematically finer-grained means to compare and distinguish their conceptual aspects.

In the present paper, the author will try to suggest a more suitable means for comparing conceptual aspects of theories of truth than other so far presented means: namely, what we will suggest is *relative truth definability*, which examine whether the truth predicate(s) of one theory Q is definable in another theory S in terms of S's truth predicate(s). Although its precise definition and detailed explanations will be given in §3, I briefly explain it here. Given two theories Q and S of truth over languages $\mathcal{L}_{\mathsf{Q}}$ and $\mathcal{L}_{\mathsf{S}}$ which share the same base theory B over $\mathcal{L}_{\mathsf{B}}$ ($\subset \mathcal{L}_{\mathsf{Q}} \cap \mathcal{L}_{\mathsf{S}}$), we say that Q is relatively truth definable in S, if for each truth predicate T of $\mathcal{L}_{\mathsf{Q}}$ there exists a predicate R_T of $\mathcal{L}_{\mathsf{S}}$ such that $\mathsf{Q} \vdash \phi$ implies $\mathsf{S} \vdash \phi'$, where ϕ' is obtained from ϕ by replacing all the occurrence of each truth predicate T by R_T; in other words, Q is relatively truth definable in S, if there exists $\mathcal{L}_{\mathsf{B}}$-conservative interpretation of Q in S. This notion of relative truth definability is in fact a finer-grained tool for comparing theories of truth than mere derivability of truth-free sentences or other well-known means. For example, conservativity for truth-free theorems and relative interpretability follow from relative truth definability, but the converse does not necessarily hold. We will then argue that relative truth definability sheds more light on conceptual aspects of theories of truth and compare various theories in the light of it. In the end, we will see that relative truth definability can distinguish some theories with the same truth-free theorems but based on apparently different views of truth and also that it is useful in its applications.

The author does not at all intend to deny other kinds of comparison of axiomatic theories of truth. What the present paper aims at is rather to suggest one possible methodology for comparison and investigation of more conceptual aspects of truth theories, to which not enough attention has been paid, and to give a new perspective of recent works of axiomatic truth theories in the light of relative truth definability. Thus, the present paper may also be viewed as a survey of this research area from a specific point of view, although it contains several new proof-theoretic results including complete answers to the conjectures raised by Feferman in [13].

§2. List of theories of truth. Davidson argued that it is impossible to define truth in terms of more basic terminologies.[1] As we abandon the attempt to define truth, truth must be conceived as primitive in our language and thus our formal framework is assumed to be provided with a primitive truth predicate(s). Then we need to give axioms or inference rules which are supposed to govern that primitive truth predicate(s). Although it is still

[1]Indeed Davidson goes further: he abandoned any 'substitute for definition' and any attempt 'to provide a brief criterion, schema, partial but leading hint, in place of a strict definition' [5, p. 320].

arguable whether to give axioms and rules to primitives is indeed to specify their meanings, axioms and rules determine their use and how they behave within the language, and these axioms and rules should embody or reflect certain specific conceptions of truth.

In the present chapter, before getting into discussing relative truth definability in §3, we will introduce various axiomatic theories of truth to be considered. Although it is not the complete list of the theories in the literature, it covers most of the major axiomatic theories of truth. Before introducing those theories, we need to make some preliminary definitions.

2.1. Preliminaries. We focus on the cases in which theories of truth share their base theory. This is because we are interested in a comparison of the conception of truth and thus we want to keep all the other conceptions fixed thereby excluding inessential factors. For example, suppose the base theory of one theory is PA and that of another theory is ZFC. Then, in most cases, the truth of the former theory can be defined in the latter solely in terms of sets without appealing to its truth predicate. In such cases, what conception of the truth is employed in the latter theory does not matter at all. Hence, we start by fixing a base theory B to which every theory is obtained by adding truth predicate(s) and its (their) axioms and inference rules.

We follow the convention and assume that the base theory B is a first-order classical system of arithmetic. We are principally interested in the case where B is Peano Arithmetic PA, but we adopt a slightly more general setting particularly in order to deal with Feferman's schematic reflective closures (see §3.3) together in the same framework. In the present paper, the language of PA, written $\mathcal{L}_{\mathsf{PA}}$, with equality as its only relation symbol and constant $\bar{0}$ for zero as its only constant symbol, is assumed to contain function symbols for each primitive recursive function and PA contains axioms defining them. Given a language $\mathcal{L}' \supset \mathcal{L}_{\mathsf{PA}}$, *full induction for $\mathcal{L}'$* is the following schema:

$$\phi(0) \land \forall x\big(\phi(x) \to \phi(x+1)\big) \to \forall x\phi(x), \text{ for each } \mathcal{L}'\text{-formula } \phi.$$

Then, our base theory B is any extension of PA such that **(a)** its language $\mathcal{L}_{\mathsf{B}}$ comprises $\mathcal{L}_{\mathsf{PA}}$ and possibly finitely many auxiliary relation symbols, **(b)** its axioms may include a primitive recursive subschema of full induction for $\mathcal{L}_{\mathsf{B}}$, **(c)** its axioms may contain at most finitely many sentences other than the specified in (b), **(d)** it is arithmetically sound.[2] We fix such a base theory B throughout the present paper and $\mathcal{L}_0$ denotes the language $\mathcal{L}_{\mathsf{B}}$ of B. All the statements in the following are valid regardless of the choice of the base theory B unless otherwise stated, subject to the conditions (a)–(d). For a given theory Q, $\mathcal{L}_{\mathsf{Q}}$ denotes the language of Q. The language of every theory of truth presented in this paper is an expansion of $\mathcal{L}_0$ with a primitive

[2]Arithmetical soundness is only needed for Lemma 32, Lemma 33 and Lemma 34 and dispensable for all the other results of the present paper. See also fn. 38, fn. 39 and fn. 40.

recursive set of auxiliary predicates. Hence, for any theory Q of truth in the present paper, the $\mathcal{L}_{\mathsf{Q}}$-terms are the same as the $\mathcal{L}_{\mathsf{PA}}$-terms.

In formulating a theory Q of truth, we have to fix some coding (Gödel numbering) of the language to which the truth predicate is to be applied. First of all, we assume that, for any language $\mathcal{L}$, the logical connectives are '$\neg$' (negation), '$\vee$' (disjunction) and '$\rightarrow$' (conditional) and the quantifier is '$\forall$' (universal quantification); the conjunction '$\wedge$' and the existential quantifier '$\exists$' are defined in the usual manner.[3]

Given a language $\mathcal{L}$, we assume that each syntactical expression e from the vocabulary of $\mathcal{L}$ is assigned a fixed Gödel number or 'code' denoted by $\#e$. Let $\ulcorner e \urcorner$ denote the term (numeral) of $\mathcal{L}$ representing $\#e$; for a precise definition of the coding, we refer the readers to [11] for example. In formulating a theory of truth, we also need to introduce the formal representations of some relations and operations which enable us to talk about the syntax of $\mathcal{L}$ within the theory; it can be seen from our assumptions on the base theory B and theories of truth that all the relations and functions (except the value function $\mathrm{val}(x)$) introduced below are primitive recursive and we assume that they are represented by means of corresponding symbols in $\mathcal{L}_0$.

For a given language $\mathcal{L}$, we first introduce some primitive recursive relations for its basic syntactical notions:

$$\mathrm{Term}_{\mathcal{L}}(x)\ \big(\mathrm{CT}_{\mathcal{L}}(x)\big) \Leftrightarrow \text{`}x\text{ is a code of a (closed) term of }\mathcal{L}\text{'};$$

$$\mathrm{Var}_{\mathcal{L}}(x) \Leftrightarrow \text{`}x\text{ is a code of a variable of }\mathcal{L}\text{'};$$

$$\mathrm{AtFml}_{\mathcal{L}}(x)\ \big(\mathrm{AtSent}_{\mathcal{L}}(x)\big) \Leftrightarrow \text{`}x\text{ is a code of an atomic formula (sentence) of }\mathcal{L}\text{'};$$

$$\mathrm{Fml}_{\mathcal{L}}(x)\big(\mathrm{Sent}_{\mathcal{L}}(x)\big) \Leftrightarrow \text{`}x\text{ is a code of a formula (sentence) of }\mathcal{L}\text{'}.$$

We will often omit the subscript $\mathcal{L}$ when it is clear from the context. In particular, since $\mathcal{L}$-terms and $\mathcal{L}$-variables are identical with those of $\mathcal{L}_{\mathsf{PA}}$, we always drop the subscript '$\mathcal{L}$' from '$\mathrm{CT}_{\mathcal{L}}$' and '$\mathrm{Var}_{\mathcal{L}}$'.

Since the base theory extends PA, each language $\mathcal{L}$ of a theory of truth has the (standard) primitive recursive representations of the following syntactical

[3]In the usual formulation of a classical first-order language, we need only to introduce the negation and disjunction (or conjunction). This is because other connectives can be defined in terms of these two and regarded as mere abbreviations; for example, '$A \rightarrow B$' is defined as '$\neg A \vee B$'. However, it is not necessarily the case in other logics such as intuitionistic logic, where '$\rightarrow$' and '$\wedge$' must be introduced as independent of '$\neg$' and '$\vee$'. In fact, in the case of the inner logic of Feferman's new theory DT [13], the conditional '$\rightarrow$' is not definable by '$\neg$' and '$\vee$'; namely, $T(\ulcorner \phi \rightarrow \psi \urcorner)$ and $T(\ulcorner \neg\phi \vee \psi \urcorner)$ are not equivalent in general. Thus, in order for us to treat DT and other systems altogether, we introduce '$\rightarrow$' separately.

operations:

$$\begin{aligned} \mathrm{num}(n) &= \ulcorner n \urcorner, & \ulcorner t \urcorner \underset{\cdot}{=} \ulcorner s \urcorner &= \ulcorner t = s \urcorner, \\ \underset{\cdot}{\neg} \ulcorner \phi \urcorner &= \ulcorner \neg\phi \urcorner, & \ulcorner \phi \urcorner \underset{\cdot}{\vee} \ulcorner \psi \urcorner &= \ulcorner \phi \vee \psi \urcorner, \\ \ulcorner \phi \urcorner \underset{\cdot}{\rightarrow} \ulcorner \psi \urcorner &= \ulcorner \phi \rightarrow \psi \urcorner, & \underset{\cdot}{\forall}(\ulcorner v_k \urcorner, \ulcorner \phi \urcorner) &= \ulcorner \forall v_k \phi \urcorner, \\ \underset{\cdot}{R}(\ulcorner t_1 \urcorner, \dots, \ulcorner t_k \urcorner) &= \ulcorner R(t_0, \dots, t_k) \urcorner, & \mathrm{sb}(\ulcorner e \urcorner, \ulcorner t \urcorner, \ulcorner v_k \urcorner) &= \ulcorner e[t/v_k] \urcorner, \end{aligned}$$

where $n \in \mathbf{N}$, t, s and $t_1, \dots, t_k$ are terms, ϕ and ψ are formulae, v_k is the k-th variable, R is a predicate symbol of $\mathcal{L}$ with the arity k, and $e[t/v_k]$ is the expression obtained from an expression e by replacing each (free) occurrence of a variable v_k by a term t. We also need the 'value function' $\mathrm{val}(x)$ such that $\mathrm{val}(\ulcorner t \urcorner) = t$ for closed terms t. This value function $\mathrm{val}(x)$ is recursive and PA-definable but not primitive recursive, since we could otherwise construct a primitive recursive evaluation function for all the n-ary primitive recursive functions for each n.[4]

For the sake of simplicity, we introduce some abbreviations. First, we write $\underset{\cdot}{\forall}\, a.b$ for $\underset{\cdot}{\forall}(a, b)$ and $a(b/c)$ for $\mathrm{sb}\big(a, \mathrm{num}(b), c\big)$; when it is clear from the context, we drop c and simply write $a(b)$. Second, for each formula $\phi(v)$, we write $\ulcorner \phi(\dot{x}) \urcorner$ for $\mathrm{sb}(\ulcorner \phi(v) \urcorner, \mathrm{num}(x), \ulcorner v \urcorner)$, i.e., $\ulcorner \phi(v) \urcorner(x/\ulcorner v \urcorner)$. This definition is extended for multi-variable cases in an obvious manner. Third, we make use of the vector notation; e.g., we write $\ulcorner \phi(\dot{\vec{x}}) \urcorner$ for $\ulcorner \phi(\dot{x}_0, \dots, \dot{x}_n) \urcorner$.

In the rest of the present section, we introduce various theories of truth to be considered later. We have fixed an arbitrary base theory B, but when we need to mention a theory with a specific base theory C we will write Q⟦C⟧ for a theory of truth Q with a base theory C: for example, TB⟦PA⟧ is a theory of Tarskian biconditionals (introduced in §2.2) with the base theory PA.

2.2. Disquotational theories of truth. The background idea of disquotational theories of truth is that truth is to be axiomatized by Tarskian *T-biconditionals*, that is, equivalences of the form

$$\ulcorner \phi \urcorner \text{ is true iff } \phi,$$

which Tarski [41, 40] regarded as the 'material adequacy' condition for the definition of truth. As is well-known, if we allow ϕ to range over arbitrary sentences that may contain the truth predicate, the resulting theory becomes inconsistent due to the Liar paradox. Hence, any disquotational approach to an axiomatization of truth must impose a certain restriction to the class of sentences over which ϕ above can range; we call such a class in a given theory the *domain* of T-biconditionals.

The most straightforward and indeed oldest disquotational theory is TB, which dates back to Tarski [41, pp. 255–257], though he himself rejected it because of its too weak deductive and expressive power.

[4]For more details of the representation of syntax within arithmetical theories, we refer the readers to [2] or [11].

Definition 1. TB is a theory over $\mathcal{L}_{\mathrm{T}} = \mathcal{L}_0 \cup \{T\}$ whose axioms consist of B-axioms and the TB-schema:

$$T(\ulcorner \sigma \urcorner) \leftrightarrow \sigma,$$

for each T-free sentence σ. Thus, the domain of T-biconditionals is restricted to T-free sentences in TB.

A natural expansion of T-biconditional is to allow parameters in ϕ: i.e., a schema of the form

$$T(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \leftrightarrow \phi(\vec{x}).$$

This form of T-biconditional is called *uniform* T-biconditional.

Definition 2. UTB is a theory over $\mathcal{L}_{\mathrm{T}}$ whose axioms consist of B-axioms and the uniform T-biconditionals for all the T-free formulae.

The next proposition is well-known.

Proposition 3. TB *and* UTB *are proof-theoretically equivalent to* B *for* $\mathcal{L}_0$.[5]

Here we use Feferman's notion of proof-theoretic reducibility and equivalence. For their precise definitions, we refer the readers to [10] or [12]; we will also discuss proof-theoretic reducibility in connection with relative truth definability in §3.2. We follow Feferman's notations for proof-theoretical reducibility and equivalence: i.e., $\mathsf{Q}_0 \leq \mathsf{Q}_1\ [\Phi]$ means that Q_0 is proof-theoretically reducible into Q_1 for a class of formulae Φ which is assumed to be a subset of $\mathcal{L}_{\mathsf{Q}_0} \cap \mathcal{L}_{\mathsf{Q}_1}$; $\mathsf{Q}_0 \equiv \mathsf{Q}_1\ [\Phi]$ means that they are proof-theoretically equivalent for Φ, i.e., $\mathsf{Q}_0 \leq \mathsf{Q}_1\ [\Phi]$ and $\mathsf{Q}_1 \leq \mathsf{Q}_0\ [\Phi]$. In the following, when considering proof-theoretical reducibility for the base language $\mathcal{L}_0$, we will simply say 'proof-theoretically reducible' or 'proof-theoretically equivalent' without mentioning $\mathcal{L}_0$ and write $\mathsf{Q}_0 \leq (\equiv) \mathsf{Q}_1$ dropping the part '$[\mathcal{L}_0]$'.

We say Q_0 is conservative over Q_1 for a set of formulae $\Phi \subset \mathcal{L}_{\mathsf{Q}_0} \cap \mathcal{L}_{\mathsf{Q}_1}$, written $\mathsf{Q}_0 \subset_{\Phi} \mathsf{Q}_1$, iff every Φ-theorem of Q_0 is derivable in Q_1 as well. Then it is immediately observed that conservativity for Φ follows from proof-theoretic reducibility for Φ. Hence, for example, it follows from Proposition 3 that $\mathsf{TB}[\![\mathsf{PA}]\!], \mathsf{UTB}[\![\mathsf{PA}]\!] \subset_{\mathcal{L}_0} \mathsf{PA}$: thus, they all have the same truth-free theorems. One more point is to be attended: if $\mathsf{Q}_0 \subset_{\mathcal{L}_0} \mathsf{Q}_1$ then the consistency of Q_1 implies that of Q_0, written $\mathsf{Q}_0 \leq_{\mathrm{Con}} \mathsf{Q}_1$.

From the standard proof of Proposition 3, we easily obtain:

Proposition 4. *A theory* Q *is said to be reflexive if* Q *proves the consistency of every finite subtheory of* Q. *Then, if* B *is reflexive, so are* TB *and* UTB.

[5]For the proof, see [22] or [25]. As a matter of fact, we can say more. Suppose B contains full induction for $\mathcal{L}_0$. Then, even if we add full induction for $\mathcal{L}_{\mathrm{T}}$ to TB or UTB, the resulting theory still remains conservative over B, since we can replace T by a definable partial truth predicate for a fixed complexity in each given derivation (see [19]). This idea is in fact what Tarski originally presented in [41]. However, since induction for partial truth predicates is not necessarily available in B in our current setting, this proof does not work for Proposition 3 in general.

The domains of T-biconditionals are restricted to arithmetical sentences in TB and UTB, but one can consider a disquotational theory of truth whose domain of T-biconditionals is larger and may contain sentences with T. However, it has turned out to be difficult to specify such a domain in a natural and consistent way among uncountably many possibilities.[6] As far as the author knows, only one such expansion of the domain has so far been presented; the schema PUTB is the uniform T-biconditionals with the domain the set of sentences in which T occurs only positively. The acronym PUTB stands for *positive uniform T-biconditionals*.[7]

The schema PUTB properly extends UTB, but if we add PUTB to B, it is known that the resulting theory is still conservative over B. On the other hand, if we add the full induction for $\mathcal{L}_\mathrm{T}$ together with PUTB, the resulting theory is no longer conservative; we call this system PUTB.

DEFINITION 5. Given a $\mathcal{L}_\mathrm{T}$-formula ϕ, ϕ is said to be T-positive if every occurrence of T in ϕ is positive. The theory PUTB over $\mathcal{L}_\mathrm{T}$ consists of B-axioms, full induction for $\mathcal{L}_\mathrm{T}$ and the schema PUTB: for each T-positive formula $\phi(\vec{x})$,

$$\forall \vec{x}\,[T(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \leftrightarrow \phi(\vec{x})].$$

Then, PUTB^- denotes the system B plus the PUTB-schema only (without full induction for $\mathcal{L}_\mathrm{T}$); PUTB^- is proof-theoretically equivalent to B as was noted.

We shall see later that PUTB is fairly strong theory.

2.3. Hierarchical theories of truth. We next turn to theories of hierarchical truth. In his seminal paper [41], Tarski provided us with a means to define truth of a formal languages by means of the so-called ‘inductive clauses’ for truth. As is desired, his definition entails T-biconditionals for the target language. However, this definition cannot be carried out within the target language itself; this is his famous undefinability theorem of truth. These facts motivate the ‘orthodox approach’ [27] toward truth which leads to the Tarskian hierarchical view of languages and their truth. According to the ‘orthodox approach’, languages form a certain hierarchy in which the truth of one language is defined in and belongs to another language at a higher level in the hierarchy. Theories of hierarchical truth is motivated by this view

[6] McGee [31] showed that there are uncountably many different and incompatible maximal consistent sets of T-biconditionals; to make the matter worse, Cieśliński [4] further showed that there are uncountably many different and incompatible maximal sets of T-biconditionals that are even *conservative over the base theory*. It is also observed by Halbach (by private communication) that any theory can be equivalently reformulated in a purely disquotational form by means of McGee’s trick [31]. Thus the strength of purely disquotational theories can range over all the possible ones.

[7] As far as the author knows, the PUTB schema was first considered by Cantini [2], and Halbach [21] used it as a formal system of truth. The appellation PUTB is due to Halbach.

and formulated so as to incorporate Tarski's 'inductive clauses' for truth of lower languages in their axioms.

Definition 6. The theory TC^- over $\mathcal{L}_\mathrm{T}$ consists of B-axioms and Tarski's 'inductive clauses' for truth:

$$\forall \vec{x} \in \mathrm{CT}\big(T(\underset{\cdot}{R}\,\vec{x}) \leftrightarrow R\big(\mathrm{val}(\vec{x})\big)\big), \text{ for each } \mathcal{L}_0\text{-atomic } R,$$
$$\forall x \in \mathrm{Sent}_{\mathcal{L}_0}\big(T(\underset{\cdot}{\neg}\,x) \leftrightarrow \neg Tx\big),$$
$$\forall x, y \in \mathrm{Sent}_{\mathcal{L}_0}\big(T(x \underset{\cdot}{\vee} y) \leftrightarrow Tx \vee Ty\big),$$
$$\forall x, y \in \mathrm{Sent}_{\mathcal{L}_0}\big(T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow \big(Tx \rightarrow Ty\big)\big),$$
$$\forall z \forall x \big[\mathrm{Sent}_{\mathcal{L}_0}(\underset{\cdot}{\forall}\, z.x) \rightarrow \big(T(\underset{\cdot}{\forall}\, z.x) \leftrightarrow \forall y Tx(y/z)\big)\big].$$

The system TC is obtained by adding full induction for $\mathcal{L}_\mathrm{T}$ to TC^-.

The acronym $\mathsf{TC}^{(-)}$ stands for Tarski's (inductive) clauses, and the above axioms are a direct axiomatization of Tarski's 'inductive clauses'. For example, the last axiom says that a sentence $\forall z\phi(z)$ is true iff $\phi(a)$ is true for each name a of an object in the domain. Note that the truth predicate characterized in this way is essentially the same thing as the so-called satisfaction class of models of arithmetic.[8]

Theorem 7 (Halbach [19]). TC^- *is proof-theoretically equivalent to* B.[9]

On the other hand, TC is no longer conservative over B; indeed TC can prove the consistency of B (and thus TC^-). In the case of $\mathsf{B} = \mathsf{PA}$, it is known that:

Theorem 8 (Feferman [6]). $\mathsf{TC}[\![\mathsf{PA}]\!]$ *is proof-theoretically equivalent to* ACA (*the theory of arithmetical comprehension*).

This match-up is established by the standard existence proof of a truth predicate of PA in ACA (see [39, §18]) and the canonical embedding of second-order quantifications of ACA in $\mathsf{TC}[\![\mathsf{PA}]\!]$ using the inductive clauses (see [18] or [22]).

Note that since TC contains full induction for $\mathcal{L}_\mathrm{T}$ and extends PA, it is reflexive (indeed, essentially reflexive) by Lemma III.3.47 of [15]. In addition, we can also see from Halbach's proof of Theorem 7 that we can uniformly find a primitive recursive function F such that any derivation of an $\mathcal{L}_0$-sentence from a finite set $\Delta \subset \mathsf{TC}^-$ can be primitive recursively transformed into a derivation of the same sentence from a finite set $F(\Delta) \subset \mathsf{B}$ (note that this says more than mere proof-theoretic reducibility). Hence, if B is reflexive then so is TC^- (but not essentially reflexive).[10]

[8] For a detailed exposition of satisfaction classes, see [24].

[9] Historically, Kotlarski, Krajewski and Lachlan [26] first gave a model-theoretic proof that $\mathsf{TC}^-[\![\mathsf{PA}]\!]$ is conservative over PA for Π^0_∞. Then, Halbach provided a finitary proof for this conservativity via cut-elimination and generalized this result to arbitrary base theories.

[10] As far as the author knows, Proposition 9 was first explicitly noticed by Martin Fischer (by private communication).

Proposition 9. TC *is reflexive, and if* B *is reflexive then so is* TC^-.

TC^- and TC present how to axiomatize Tarski's 'inductive clauses'. Then, we can apply this to ascend up to higher and higher languages in the hierarchy of 'orthodox approach'. Recall that the hierarchical view claims that any language cannot contain its own truth predicate and the truth predicate of a given language $\mathcal{L}$ is only contained and defined in a richer language $\mathcal{L}'$ than $\mathcal{L}$; then we iterate this process of ascending to the richer and richer languages and theories from the base theory B and its language $\mathcal{L}_0$. The theories thus obtained form a hierarchy $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_m, \dots$. This hierarchy can go up beyond transfinite levels by taking the union of the lower languages at a limit stage. To formalize this idea, we first fix a standard notation system up to the least critical ordinal Γ_0;[11] we can of course consider higher levels, but Γ_0 is sufficient for our subsequent arguments. For the simplicity, we identify ordinals and their representations; e.g., $\bar{\alpha}$ denotes the numeral whose value is α in this notation system; but we often abuse notation and write α for $\bar{\alpha}$.

Definition 10. Let $\alpha \leq \Gamma_0$ and $\mathcal{L}_\alpha$ be a language $\mathcal{L}_0 \cup \{T_\beta\}_{\beta<\alpha}$ where T_β is a new unary predicate expressing the truth of β-th level. Then, the ramified truth system up to α, written as $\mathsf{RT}_{<\alpha}$, is a theory over $\mathcal{L}_\alpha$ consisting of B-axioms, full-induction for $\mathcal{L}_\alpha$ and, for each $\gamma < \beta < \alpha$,

R1: $\forall \vec{x} \in \mathrm{CT}\big(T_\beta(\underset{\cdot}{R}\,\vec{x}) \leftrightarrow R(\mathrm{val}(\vec{x}))\big)$, for each $\mathcal{L}_0$-atomic R,

R2: $\forall x \in \mathrm{Sent}_{\bar{\beta}}\big(T_\beta(\underset{\cdot}{\neg}\, x) \leftrightarrow \neg T_\beta(x)\big)$,

R3: $\forall x, y \in \mathrm{Sent}_{\bar{\beta}}\big[\big(T_\beta(x \underset{\cdot}{\vee} y) \leftrightarrow T_\beta(x) \vee T_\beta(y)\big) \wedge \big(T_\beta(x \underset{\cdot}{\rightarrow} y) \leftrightarrow T_\beta(x) \rightarrow T_\beta(y)\big)\big]$,

R4: $\forall z \forall x\big[\mathrm{Sent}_{\bar{\beta}}(\underset{\cdot}{\forall}\, z.x) \rightarrow \big(T_\beta(\underset{\cdot}{\forall}\, z.x) \leftrightarrow \forall y T_\beta x(y/z)\big)\big]$,

R5: $\forall x \in \mathrm{CT}\big[\mathrm{Sent}_{\bar{\gamma}}(\mathrm{val}(x)) \rightarrow \big(T_\beta(\underset{\cdot}{T}_\gamma\, x) \leftrightarrow T_\gamma \mathrm{val}(x)\big)\big]$,

R6: $\forall x \in \mathrm{CT} \forall \delta < \bar{\beta}\big[\mathrm{Sent}_\delta(\mathrm{val}(x)) \rightarrow \big(T_\beta(\underset{\cdot}{T}_\delta\, x) \leftrightarrow T_\beta \mathrm{val}(x)\big)\big]$,

where $\mathrm{Sent}_a(b)$ is a binary primitive recursive predicate on a and b such that $\mathrm{Sent}_a(b)$ iff $\mathrm{Sent}_{\mathcal{L}_\delta}(b)$ for $\delta = a$ and $\underset{\cdot}{T}_a\, b$ is a binary primitive recursive function on a and b whose value is $\ulcorner T_\beta t \urcorner$ when $a = \beta$ and $b = \ulcorner t \urcorner$.[12]

Then, the hierarchy of languages is represented by $\mathsf{RT}_{<\alpha}$ in the following way: a predicate T_β in $\mathcal{L}_{\beta+1}$ is the truth predicate of $\mathcal{L}_\beta$ in the sense that T_β satisfies the Tarski's inductive clauses **R1**–**R5** for $\mathcal{L}_\beta$; **R5** and **R6** express that the higher truth predicates include all the lower truth predicates and thus the higher languages are richer than the lower ones in the hierarchy; **R6** also indicates that T_λ for a limit λ is the union of all the lower truth predicates.

Suppose $\alpha < \beta$. Then, $\mathsf{RT}_{<\beta}$ can prove the consistency of $\mathsf{RT}_{<\alpha}$, since $\mathsf{RT}_{<\beta} \vdash \forall x \in \mathrm{Sent}_\alpha[\mathrm{Prv}_{\mathsf{RT}_{<\alpha}}(x) \rightarrow T_\alpha(x)]$. Hence, we have:

Proposition 11. *For* $\alpha < \beta$, $\mathsf{RT}_{<\alpha} \subsetneq_{\mathcal{L}_0} \mathsf{RT}_{<\beta}$.

[11] For its precise definition, see [8] or [34].

[12] We follow the formulation of $\mathsf{RT}_{<\alpha}$ by Halbach in [22].

Let $\mathsf{RA}_{<\alpha}$ denote the ramified analysis up to α.[13] Then the ramified truth theory and ramified analysis are related in the following way:[14]

THEOREM 12 (Feferman [6]). $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\alpha}$ *for each ordinal* α.

For $\alpha < \Gamma_0$ and a formula $\phi(x)$ of a language $\mathcal{L}$, let $\mathrm{TI}(\alpha; \phi)$ denote

$$\forall x\big(\forall y < x\phi(y) \to \phi(x)\big) \to (\forall x \leq \alpha)\phi(x),$$

and $\mathrm{TI}_{\mathcal{L}}(\alpha)$ denote the schema $\mathrm{TI}(\alpha; \phi)$ for all $\phi \in \mathcal{L}$. Then, for $\alpha \leq \Gamma_0$, $\mathrm{TI}_{\mathcal{L}}(< \alpha)$ denotes transfinite induction up to α: i.e., $\mathrm{TI}_{\mathcal{L}}(\beta)$ for all $\beta < \alpha$. It is known that, for an epsilon number α (i.e., $\alpha = \omega^\alpha$), $\mathsf{RA}_{<\alpha}$ proves transfinite induction up to $\varphi_\alpha 0$ for arithmetical formulae, i.e., $\mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(< \varphi_\alpha 0)$, where φ is the binary Veblen function. It follows from the equivalence between $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!]$ and $\mathsf{RA}_{<\alpha}$ that $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!]$ proves the same transfinite induction (cf. [8, 11]). Then, the proof of $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!] \vdash \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(< \varphi_\alpha 0)$ induced from the proof of $\mathsf{RA}_{<\alpha} \vdash \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(< \varphi_\alpha 0)$ can be straightforwardly transformed to the proof of $\mathsf{RT}_{<\alpha}[\![\mathsf{B}]\!] \vdash \mathrm{TI}_{\mathcal{L}_{\mathsf{B}}}(< \varphi_\alpha 0)$ for an arbitrary base theory B which meets the conditions we have posed. Thus, we have the next theorem.

THEOREM 13. *Let* α *be an epsilon number. Then,* $\mathsf{RT}_{<\alpha} \vdash \mathrm{TI}_{\mathcal{L}_0}(< \varphi_\alpha 0)$.

2.4. Iterative compositional theories of truth. The iterative compositional conception of truth is best represented by the following passage of Feferman [11, p. 18]: 'truth or falsity is *grounded* in atomic facts from the base language $\mathcal{L}$, i.e., can be determined from such facts by evaluation according to the rules of truth for the connectives and quantifiers, and where statements of the form $[T(\ulcorner A \urcorner)]$ are evaluated to be true (false) only when A itself has already been verified (falsified).' It is 'iterative' because a sentence $T(\ulcorner A \urcorner)$ is true only if A is already true, and it is 'compositional' because a compositional sentence is determined only by its components according to the evaluation rule for connectives and quantifiers. Kripke [27] first focused on this iterative compositional conception of truth and gave a model-theoretic formulation of such truth, and then Feferman [11] later gave an axiomatic formalization KF of this conception.[15] For simplicity, we write $F(x)$ for $T(\dot{\neg} x)$ (its intended meaning is 'false') in the following.

DEFINITION 14. The system KF over $\mathcal{L}_\mathrm{T}$ consists of B, full induction for $\mathcal{L}_\mathrm{T}$ and:

K0: $\forall z \in \mathrm{Var} \forall y \in \mathrm{CT} \forall x \big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(\dot{\forall} z.x) \to \big(T\mathrm{sb}(x, y, z) \leftrightarrow T\mathrm{sb}(x, \mathrm{num} \circ \mathrm{val}(y), z)\big)\big]$,

[13] For the precise definition, see [7] or [11].

[14] Feferman gave a sketch of its proof in [6]. A more detailed proof can be found in [18].

[15] The system KF and the main results about it were in fact presented by Feferman in 1979. For more history, see [16].

K1: for each atomic R of $\mathcal{L}_0$,

$$\forall \vec{x} \in \mathrm{CT}\big[T\big(\underset{\cdot}{R}(\vec{x})\big) \leftrightarrow R\big(\mathrm{val}(\vec{x})\big)\big] \wedge \forall \vec{x} \in \mathrm{CT}\big[F\big(\underset{\cdot}{R}(\vec{x})\big) \leftrightarrow \neg R\big(\mathrm{val}(\vec{x})\big)\big],$$

K2: $\forall x \in \mathrm{CT}\big[T(\underset{\cdot}{T}\, x) \leftrightarrow T\big(\mathrm{val}(x)\big)\big] \wedge \forall x \in \mathrm{CT}\big[F(\underset{\cdot}{T}\, x) \leftrightarrow F\big(\mathrm{val}(x)\big)\big]$,

K3: $\forall x \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[T(\underset{\cdot}{\neg}\underset{\cdot}{\neg}\, x) \leftrightarrow Tx\big]$,

K4: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[T(x \underset{\cdot}{\vee} y) \leftrightarrow (Tx \vee Ty)\big]$,

K5: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[F(x \underset{\cdot}{\vee} y) \leftrightarrow (Fx \wedge Fy)\big]$,

K6: $\forall z \forall x\big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(\underset{\cdot}{\forall} z.x) \rightarrow \big(T(\underset{\cdot}{\forall} z.x) \leftrightarrow \forall y Tx(y/z)\big)\big]$,

K7: $\forall z \forall x\big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(\underset{\cdot}{\forall} z.x) \rightarrow \big(F(\underset{\cdot}{\forall} z.x) \leftrightarrow \exists y Fx(y/z)\big)\big]$,

K8: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}} \rightarrow \big[\big(T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow T(\underset{\cdot}{\neg}\, x \underset{\cdot}{\vee} y)\big) \wedge \big(F(x \underset{\cdot}{\rightarrow} y) \leftrightarrow F(\underset{\cdot}{\neg}\, x \underset{\cdot}{\vee} y)\big)\big]$.[16]

Then, KF^- is B plus **K0**–**K8** (without adding full induction).

The axioms of KF is regarded as a partial axiomatization of (closed off) Kripkean fixed-point model with Strong Kleene schema. As is known, Strong Kleene Logic is the three valued logic with the following truth table:

$\neg$	
T	F
U	U
F	T

$\vee$	T	U	F
T	T	T	T
U	T	U	U
F	T	U	F

The conditional '$\rightarrow$' and conjunction '$\wedge$' are defined in the usual manner in terms of '$\neg$' and '$\vee$'. The existential and universal quantifiers are defined as infinitary disjunction and conjunction respectively.

Now it is clear how KF represents the iterative compositional conception of truth with Strong Kleene schema. For example, **K5** represents how falsity is assigned to a disjunctive sentence in Strong Kleene schema: that is, a disjunctive sentence $\phi \vee \psi$ is false iff both conjuncts are false. For another example, **K2** represents the iterative character of truth in which a sentence $T(\phi)$ is true (false) iff ϕ is true (false). The axiom **K0** is postulated to assure that the truth of a given sentence $\phi(t)$ depends only on the value of t.[17]

The proof-theoretic strength of KF and KF^- are known:

THEOREM 15 (Feferman [11]). $\mathsf{KF}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\varepsilon_0}$. *Thus* $\mathsf{KF}[\![\mathsf{PA}]\!]$ *is also equivalent to, e.g.,* Σ^1_1-AC *and* $\widehat{\mathsf{ID}}_1$.[18]

THEOREM 16. $\mathsf{KF}^- \equiv \mathsf{B}$.

[16] This axiom is not usually postulated in the literature, but we need it since we have assumed that '$\rightarrow$' is a primitive logical connective and since '$\rightarrow$' is definable by '$\neg$' and '$\vee$' in Strong Kleene Logic.

[17] **K0** was dropped in Feferman's original formulation of KF, since it is derivable from the other axioms in KF. However, it is not derivable in KF^- without full induction for $\mathcal{L}_\mathrm{T}$ and thus introduced by Cantini for KF^-. It is sometimes called the regularity axiom.

[18] For the precise definitions of Σ^1_1-AC and $\widehat{\mathsf{ID}}_1$, see [9].

The equivalence $\mathsf{KF}^-[\![\mathsf{PA}]\!] \equiv \mathsf{PA}$ is observed from Cantini's proof of Theorem 9.10 of [2], where he showed that $\mathsf{KF}_{\mathrm{tot}}$, an extension of $\mathsf{KF}^-[\![\mathsf{PA}]\!]$, is reducible into PA. $\mathsf{KF}_{\mathrm{tot}}$ is not reflexive and indeed Cantini's proof does not directly yield the reflexivity of $\mathsf{KF}^-[\![\mathsf{PA}]\!]$, since the bound of complexity of the reduced proof in PA depends on the length of the original proof in $\mathsf{KF}_{\mathrm{tot}}$. However, Halbach's technique in [19] is applicable to show $\mathsf{KF}^- \equiv \mathsf{B}$ for an arbitrary base theory B, and thus yields the reflexivity result in the same manner as Proposition 9.

PROPOSITION 17. KF *is* (*essentially*) *reflexive. If* B *is reflexive, then so is* KF^-.

It is also known that PUTB is a subtheory of KF; more precisely, we have:

LEMMA 18 (Cantini [2]). KF^- *proves the* PUTB *schema. Hence* PUTB^- (PUTB) *is a subtheory of* KF^- (KF *respectively*).

In the context of iterative compositional theories of truth, we sometimes consider the two special axioms: the consistency axiom and completeness axiom which are defined respectively by

$$\text{Cons: } \forall x \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}} \neg(Tx \wedge Fx) \qquad \text{Comp: } \forall x \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(Tx \vee Fx).$$

Feferman [11] argued that the character of Cons is different from the rules of evaluation incorporated in KF and his argument should apply to Comp as well.[19] In fact, neither of them is derivable from KF and adding both to KF leads to a contradiction.[20] On the other hand, Cantini [2] showed that either (but not both) of them can be consistently and even conservatively (for $\mathcal{L}_0$) added to KF. We will see in §2.5 and §5 that Cons is particularly important in the connection between iterative compositional theories and determinate theories of truth.

Concluding this subsection, we have a closer look at iterativity and compositionality. We can now extract from the above observations more general characteristics of the iterative compositional theories of truth. That is, they may be characterized in the following way: for each logical connectives C, its axioms for compositional truth are of the form

$$\forall \vec{x}\big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(x_0) \wedge \cdots \wedge \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(x_k) \to \big(T(\underset{\cdot}{C}\,\vec{x}) \leftrightarrow R_C(\vec{x})\big)\big] \qquad (1)$$

for a certain formula R_C suitable for a given evaluation rule for C; the axioms for quantifiers are of the parallel forms for each quantifier Q and its evaluation formula R_Q; the axioms for iterativity are of the form like

[19] Feferman denotes Cons by *Disj* there. The notations Cons and Comp are due to Cantini [2].

[20] Given a least Kripkean fixed-point model M with Strong Kleene schema, if M is closed off in the way that all the undetermined sentences are pushed into the *extension* of T, then the resulting classical structure is a model of KF + Comp; if M is closed-off in the usual manner then it is a model of KF + Cons.

$T(TA) \leftrightarrow TA$ as in KF. Although it is highly debatable what kinds of evaluation formulae R_C and R_Q can be regarded as 'compositional', any system for iterative compositional truth should be axiomatized in this way. Now, let S be an iterative compositional theory of truth with the full induction, and suppose that R_C and R_Q are T-positive for each connective C and quantifier Q. Then, we can show that the truth of S can be constructed as a fixed-point of a positive arithmetical operator. Therefore, when the base theory B is PA, S is proof-theoretically reducible to Σ^1_1-AC via Aczel's construction of a Σ^1_1 fixed-point of a Σ^1_1-formula (see [11]).

2.5. Determinate theory of truth. Feferman's theory DT [13] brought about a new form of axiomatic theories of truth. The background idea of DT, which Feferman ascribes to Russell, is that every predicate R has a domain D of significance and it makes sense to apply R only to objects in D and therefore the principles which are supposed to characterize the concept expressed by R are to be applied only to objects in D. Thus the theory of truth based on this view would have two predicate T and D representing truth and its domain of significance. According to Feferman, such a domain of significance in the case of truth consists of *meaningful and determinate* sentences, i.e., sentences with a definite truth value, true or false, and the T-schema and other principles characterizing truth can be restricted to such sentences. Feferman's argument for this identification of the domain with the set of determinate sentences is simple. Let D be the domain of truth predicate T. Then, we have $[T(A) \vee F(A)] \rightarrow D(A)$.[21] Conversely, since the T-biconditional, $T(A) \leftrightarrow A$ and $F(A) \leftrightarrow \neg A$, obtains for every A in D, we have $D(A) \rightarrow [T(A) \vee F(A) \leftrightarrow (A \vee \neg A)]$; thus, we have $D(A) \rightarrow [T(A) \vee F(A)]$. This argument is intuitively convincing and Feferman just identifies $D(x)$ with $T(x) \vee F(x)$; thus the new predicate D becomes redundant.[22] In addition, Feferman further requires that D must be *strongly compositional* in the sense that a compound sentence is in D iff all the substitution instances of its subformulae by meaningful terms belong to D.[23] Feferman presented a system DT to capture these views on truth.

[21]To be precise, in this argument, Feferman implicitly assumes that $D(\neg A) \rightarrow D(A)$.

[22]Formally speaking, there is a subtle matter about this argument. The argument to conclude $D(A) \leftrightarrow [T(A) \vee F(A)]$ for each (standard) sentence A does not formally entail $D(x) \leftrightarrow [T(x) \vee F(x)]$ for all x. This is because any (recursive) arithmetical theory ($\supset$ PA) must have a nonstandard model M and thus nonstandard sentences in M by the usual overspill argument. Hence, when one introduces a predicate D separately from T, the equivalence $D(x) \leftrightarrow T(x) \vee F(x)$ is not necessarily *provable*. It may be of some interest to consider and investigate a DT-like theory in which D and T are separately introduced.

[23]As Feferman himself pointed out, strong compositionality in this sense is not met with respect to the conditional '$\rightarrow$' due to **D11**. If we pose full strong compositionality, the resulting theory becomes identical to the theory WKF + Cons which will be introduced later. See §5.

DEFINITION 19. DT^- is a theory over $\mathcal{L}_\mathrm{T}$ whose axioms consist of B-axioms, **K0** and:

D1: $\forall x[\mathrm{AtSent}_{\mathcal{L}_0}(x) \to D(x)]$,
D2: $\forall x \in \mathrm{CT}\big[D\big(\underset{\cdot}{T}\, x\big) \leftrightarrow D\big(\mathrm{val}(x)\big)\big]$,
D3: $\forall x \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[D(\underset{\cdot}{\neg}\, x) \leftrightarrow D(x)\big]$,
D4: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[D(x \underset{\cdot}{\vee} y) \leftrightarrow (Dx \wedge Dy)\big]$,
D5: $\forall z \forall x\big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(\underset{\cdot}{\forall}\, z.x) \to \big(D(\underset{\cdot}{\forall}\, z.x) \leftrightarrow \forall y D x(y/z)\big)\big]$,
D6: $\forall x \in \mathrm{CT}\big[T(\underset{\cdot}{R}\, \vec{x}) \leftrightarrow R\big(\mathrm{val}(\vec{x})\big)\big]$, for each atomic R of $\mathcal{L}_0$,
D7: $\forall x \in \mathrm{CT}\big[D\big(\mathrm{val}(x)\big) \to \big(T(\underset{\cdot}{T}\, x) \leftrightarrow T(\mathrm{val}(x))\big)\big]$,
D8: $\forall x \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[D(x) \to \big(T(\underset{\cdot}{\neg}\, x) \leftrightarrow \neg Tx\big)\big]$,
D9: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[D(x \underset{\cdot}{\vee} y) \to \big(T(x \underset{\cdot}{\vee} y) \leftrightarrow (Tx \vee Ty)\big)\big]$,
D10: $\forall z \forall x\big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(\underset{\cdot}{\forall}\, z.x) \wedge D(\underset{\cdot}{\forall}\, z.x) \to \big(T(\underset{\cdot}{\forall}\, z.x) \leftrightarrow \forall y T x(y/z)\big)\big]$,
D11: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[D(x \underset{\cdot}{\to} y) \leftrightarrow \big(Dx \wedge (Tx \to Dy)\big)\big]$,
D12: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[D(x \underset{\cdot}{\to} y) \to \big(T(x \underset{\cdot}{\to} y) \leftrightarrow (Tx \to Ty)\big)\big]$,

where $D(x)$ abbreviates $Tx \vee Fx$. Then, DT is DT^- plus full induction for $\mathcal{L}_\mathrm{T}$.

As is desired, it is shown by straightforward meta-induction on $\phi(\vec{x})$ that

$$\mathsf{DT}^- \vdash \forall \vec{x}\big(D(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \to (T(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \leftrightarrow \phi(\vec{x}))\big).$$

DT-axioms look different from KF-axioms *prima facie* and they have distinct background motivations. However, DT and KF are in fact very closely related. In order to see this, we first think of the 'inner logic' behind DT.

In his consistency proof of DT, Feferman constructed a Kripkean fixed-point model for DT in which the truth of sentences are evaluated according to a special logic. Let us call it Feferman Logic and this is the 'inner logic' of DT.

Feferman Logic: Feferman Logic is a variant of Weak Kleene logic with a special evaluation for the conditional '$\to$'. In Feferman Logic, the conditional '$\to$' cannot be defined in terms of '$\neg$' and '$\vee$' and the evaluation of '$\to$' is determined by a truth table independent of those for '$\neg$' and '$\vee$'; indeed $T(\underset{\cdot}{\neg}\, x \underset{\cdot}{\vee} y)$ and $T(x \underset{\cdot}{\to} y)$ are not equivalent in DT. The truth table for logical connectives are given by the following:

$\neg$	
T	F
U	U
F	T

$\vee$	T	U	F
T	T	U	T
U	U	U	U
F	T	U	F

$\to$	T	U	F
T	T	U	F
U	U	U	U
F	T	T	T

The universal quantifier is interpreted as infinitary (Weak Kleene) conjunction: i.e., $\forall x \phi(x)$ is true iff $\phi(x)$ is true for all x, and $\forall x \phi$ is false iff $\phi(x)$ is false for some x and $\phi(x)$ is either true or false for all x. The other logical primitives are defined in the usual manner by '$\neg$', '$\vee$' and '$\forall$'.

We can formulate an iterative compositional theory FKF of truth whose truth predicate is governed by this new evaluation schema of Feferman Logic. FKF is a variant of KF in which the axioms for compositionality are given according to not Strong Kleene Logic but Feferman Logic.

DEFINITION 20. FKF^- over $\mathcal{L}_\mathrm{T}$ consists of all the axioms of B and **K0**–**K3**, **K5**–**K6** (without **K8**) and:

FK4: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[T(x \mathbin{\dot\vee} y) \leftrightarrow \big((Tx \wedge Ty) \vee (Fx \wedge Ty) \vee (Tx \wedge Fy)\big)\big]$,

FK7: $\forall z \forall x \Big[\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(\dot\forall z.x) \rightarrow \Big(F(\dot\forall z.x) \leftrightarrow \big(\forall y\big(Tx(y/z) \vee Fx(y/z)\big) \wedge \exists y Fx(y/z)\big)\Big)\Big]$,

FK9: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[T(x \mathbin{\dot\rightarrow} y) \leftrightarrow \big((Tx \wedge Ty) \vee Fx\big)\big]$,

FK10: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}\big[F(x \mathbin{\dot\rightarrow} y) \leftrightarrow (Tx \wedge Fy)\big]$.

Then FKF is defined as FKF^- plus full induction for $\mathcal{L}_\mathrm{T}$.

Observe that these axioms represent how compound sentences are evaluated according to Feferman Logic in the form of (1) in §2.4; for example, **FK4** says that a compound sentence $A \vee B$ is evaluated to be true when both A and B has a definite truth value and either of them is true; **FK9** and **FK10** express that the conditional '$A \rightarrow B$' is not mere abbreviation of '$\neg A \vee B$' and evaluated according to the above truth table of Feferman Logic.

Then the next theorem indicates that determinate theories and iterative compositional theories bear a certain very close connection (we will see another example in §5).

THEOREM 21. DT (DT^-) *and* $\mathsf{FKF} + \mathrm{Cons}$ $(\mathsf{FKF}^- + \mathrm{Cons})$ *are identical theories.*

PROOF. The proof is straightforwardly done by showing all the axioms of one are provable in the other. We will only illustrate two typical examples. For $x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}$, if $Tx \wedge Fx$ then Dx and thus $Fx \leftrightarrow \neg Tx$ in DT; thus $\mathsf{DT} \vdash \mathrm{Cons}$. For another example, in $\mathsf{FKF} + \mathrm{Cons}$, given any $x, y \in \mathrm{Sent}_{\mathcal{L}}$, if $T(x \mathbin{\dot\rightarrow} y)$ then $(Tx \wedge Ty) \vee Fx$ by **FK9**. But it follows from Cons that $(Tx \wedge Ty) \vee \neg Tx$; we have $\neg Tx \vee Ty$. Conversely, if $\neg Tx \vee Ty$ and $D(x \mathbin{\dot\rightarrow} y)$, then $\neg Tx \vee (Tx \wedge Ty)$ and Dx. Since $\neg Tx$ implies Fx by Dx, we have $(Tx \wedge Ty) \vee Fx$. Finally, $\mathsf{FKF} + \mathrm{Cons} \vdash$ **D11**. ⊣

Feferman [13] raised a conjecture about the proof-theoretic strength of DT.

FEFERMAN'S FIRST CONJECTURE. $\mathsf{DT}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\varepsilon_0}$.

Feferman also suggested its proof. As far as the author sees, his suggested proof does work for the lower bound of $\mathsf{DT}[\![\mathsf{PA}]\!]$, but not for the upper bound as it is. He suggested to simulate the construction of Kripkean fixed-point model of $\mathsf{DT}[\![\mathsf{PA}]\!]$, which is given in his consistency proof of $\mathsf{DT}[\![\mathsf{PA}]\!]$, within Σ^1_1-AC, as in the proof for the upper bound of $\mathsf{KF}[\![\mathsf{PA}]\!]$ in [11]. However, as we have seen, every fixed-point model of DT must be consistent in the sense that $Ta \wedge Fa$ for no a, and Σ^1_1-AC only assures the existence of a fixed-point of an Σ^1_1 positive operator but not necessarily consistent one. In §5, we will give a solution to this obstacle and positive answer to the conjecture together with his second conjecture explained below as one application of relative truth definability.

2.6. Complete and consistent theories of truth. Theories like KF and DT are called type-free theories of truth, since they allow us to prove the truth of sentences which contain the truth predicate itself. However, neither of them can be consistently closed under Cons and Comp. Inconsistency is easily derived from Cons and Comp together with some iterative and compositional axioms.[24] In the present section, we introduce a type-free theory FS which derives both Cons and Comp. The acronym FS comes from Friedman and Sheard who first introduced this theory [14]. This appellation and the following formulation is due to Halbach [17].

Definition 22. A theory FS over $\mathcal{L}_{\mathrm{T}}$ consists of B-axioms, full induction for $\mathcal{L}_{\mathrm{T}}$, the following axioms

FS1: $\forall \vec{x} \in \mathrm{CT} \to [T(\underset{\cdot}{R}\,\vec{x}) \leftrightarrow R(\mathrm{val}(\vec{x}))]$,
FS2: $\forall x \in \mathrm{Sent}_{\mathcal{L}_{\mathrm{T}}}[T(\underset{\cdot}{\neg}\, x) \leftrightarrow \neg Tx]$,
FS3: $\forall x, y \in \mathrm{Sent}_{\mathcal{L}_{\mathrm{T}}}\big[\big(T(x \underset{\cdot}{\vee} y) \leftrightarrow (Tx \vee Ty)\big) \wedge \big(T(x \underset{\cdot}{\to} y) \leftrightarrow (Tx \to Ty)\big)\big]$,
FS4: $\forall z \forall x\big[\mathrm{Sent}_{\mathcal{L}_{\mathrm{T}}}(\underset{\cdot}{\forall}\, z.x) \to \big(T(\underset{\cdot}{\forall}\, z.x) \leftrightarrow \forall y Tx(y/z)\big)\big]$,

and the following two extra inference rules: for each $\mathcal{L}_{\mathrm{T}}$-sentence σ,

$$\frac{\sigma}{T(\ulcorner \sigma \urcorner)}\ (\mathrm{NEC}) \qquad\qquad \frac{T(\ulcorner \sigma \urcorner)}{\sigma}\ (\mathrm{CONEC})$$

The proof-theoretic strength of FS was determined by Halbach [17]. FS seems natural and has desirable properties, i.e., Cons and Comp, but it turned out to be ω-inconsistent as can be shown by using McGee's trick [30].

Theorem 23. (1) $\mathsf{FS} \equiv \mathsf{RT}_{<\omega}$. (2) FS *is* ω*-inconsistent.*

Hence, if the base theory is PA, then we have $\mathsf{FS}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\omega}$. Rathjen [35] showed that ACA_0 + BR, where BR denotes the Bar Rule, and ACA_0^+, i.e., ACA_0 plus the axiom of the existence of ω-th Turing Jump of each set X,[25] prove the same Π^1_1 theorems. Then, we can observe from Rathjen's proof that $\mathsf{RA}_{<\omega}$ (and thus $\mathsf{FS}[\![\mathsf{PA}]\!]$) also has the same arithmetical theorem as they have.[26]

[24] KF and DT cannot be consistently closed under NEC and CONEC either.

[25] For the precise definition and more explanations of ACA_0^+, see [38, Ch.X].

[26] It is noted that Rathjen, in fact, already showed that $\mathsf{RA}_{<\omega} \leq \mathsf{ACA}_0$ + BR $[\Pi^0_\infty]$ in [35]. The converse (ACA_0 + BR $\subset_{\Pi^0_\infty}$ $\mathsf{RA}_{<\omega}$) is obtained by observing that the system ACA_0^ω, which Rathjen intermediately uses to establish his result, is embeddable in $\mathsf{RA}_{<\omega}$. Although the proof of the latter is model-theoretic, Theorem 54 below gives the proof-theoretical one,

There are of course other important theories of truth which we do not discuss in the present paper such as Cantini's VF (see [3]).[27] However, we restrict our attention to the theories introduced so far. In the rest of the present paper, we will consider and compare these theories in the light of relative truth definability.

§3. Truth-definability.

3.1. Comparing conceptual aspects of axiomatic truth. Now we have seen various theories with various views on truth. Some theories of truth have the same truth-free theorems, even if they apparently embody different views of truth. Those coincidences of truth-free theorems might indeed result from a certain close and intimate connection between two theories, but it seems that mere coincidences of truth-free theorems cannot solely imply the existence of such a connection. For example, let T be a theory over $\mathcal{L}_{\mathrm{T}}$ defined as B plus a single additional axiom '$T(\ulcorner 0 = 0 \urcorner)$'. Then, we can show that T and UTB are proof-theoretically equivalent (and in fact mutually interpretable if B is reflexive). However, it seems implausible to think that T and UTB share a common view of truth or they are conceptually 'equivalent'. Hence, proof-theoretic equivalence, conservativity or other well-known intertheoretic relations seem too coarse-grained for our purpose of comparing conceptual aspects of theories of truth.

We are thus interested in a finer-grained means to compare conceptual aspects of theories of truth. In the present section, we will propose such a criterion, by means of which we can represent a certain 'equivalence' or 'reducibility' among theories of truth from a more conceptual point of view, although it may be still too coarse to fully represent 'conceptual equivalence' or 'conceptual reducibility' (in some strong sense). As we have argued, any relation among theories which only focuses on truth-free parts seems insufficient for our purpose on the one hand. On the other hand, some relations are too strong. For example, subtheory relation and its induced equivalence relation, i.e., identity (as theories), should be too strong; e.g., $\mathsf{KF} + \neg T(\mathbf{n})$ for some numeral $\mathbf{n}$ which codes no sentence is not a subtheory of KF, but we usually think that the difference only consists in an inessential point irrelevant to their views of truth.[28]

One probably comes to relative interpretability as a candidate of such an intertheoretic relation. Indeed, we often say, for instance, that geometry (or any other subject of mathematics) is reducible to set theory, and this usually

since $\mathsf{ACA}_0 + \mathsf{BR}$ can be embedded in $\mathsf{TC}^*[\![\mathsf{PA}]\!]$ ($\equiv \mathsf{RA}_{<\omega}$) in the standard manner (the second-order variables are interpreted to range over $\mathrm{Sent}_{\mathcal{L}_0(P)}$).

[27] VF should be classified as 'iterative *non*-compositional' theories of truth. Cantini [3] showed that $\mathsf{VF} \equiv \mathsf{ID}_1$: that is, VF has an impredicative proof-theoretic strength.

[28] Subtheory relation as a representation of intertheoretic 'reducibility' was also criticized by Niebergall [33]. His argument is about 'theory reducibility' in general and thus is applied to our current discussion as well.

means the existence of a relative interpretation of geometry in set theory. However, it seems still arguable whether this implies that the concept of geometrical objects such as 'line' or 'triangle' are 'reducible' to the concept of sets. It is a deep and difficult problem whether this is a genuine reduction or not, but, at any rate, relative interpretability is not what we want to propose.[29]

First of all, we note that the theory T above and UTB are mutually interpretable. Hence relative interpretability cannot exclude this kind of pathological cases. In general, the problems of relative interpretability as our means of comparison are:

(A) Relative interpretation may interpret terms and predicates by totally different notions;
(B) Relative interpretation may change the range of quantifiers, in other words, the ontological commitment of the interpreted theory;
(C) Relative interpretability does not necessarily preserve truth-free theorems;
(D) Relative interpretability just follows from conservativity for truth-free theorems in some typical setting of theories of truth.

We will explain each of these problems below.

The characteristics (A) and (B) of relative interpretability are principal defects for our current objectives. It is true that relative interpretability is very useful and widely applicable for comparing diverse kinds of mathematical theories, but we are currently in a special situation in which only theories of truth are considered and we compare only their conceptions of truth. For example, relative interpretability is often regarded as a suitable tool for ontological reduction of mathematical (or other) objects posited by one mathematical subject into those posited by another subject. However, what we are aiming at is not ontological reduction but rather, so to say, 'conceptual reduction'. Thus, for our current purpose, we would like to keep every other concept fixed in order to exclude inessential factors, although it may make our approach less versatile than relative interpretability. Therefore, since relative interpretability may require that the words, e.g., 'for all natural numbers' or 'greater than' of one theory be translated by 'for all even numbers' or 'less than' in another theory, relative interpretability seems unsuitable for our current purpose.

The problem (C) is caused by (A) and (B) and it is problematic since we are looking for a finer-grained means to compare theories. There are indeed two theories Q_0 and Q_1 such that Q_1 can prove properly more truth-free theorems than Q_0 but is relatively interpretable in Q_0; e.g., consider TC and $\mathsf{TC} + \neg\mathrm{Cons}(\mathsf{TC})$. If the truth of Q_1 is 'reducible' to Q_0, then it is natural to

[29]For more discussion of relative interpretability as a tool of reduction in more general settings, particularly on its role in foundational issues, see e.g., [12] and [33].

require that the truth of Q_1 should not bring richer truth-free consequences than that of Q_0 does.

The fourth problem (D) is more precisely elaborated by the following fact.

THEOREM 24. *Let* Q *and* S *be first-order classical systems and primitive recursive extensions of* PA. *Suppose that* Q *and* S *are both reflexive and that* $\mathsf{Q} \subset_{\Pi^0_1} \mathsf{S}$. *Then* Q *is interpretable in* S.

PROOF. Let $\sigma(x)$ be a primitive recursive predicate which binumerates Q-axioms. Since Q is reflexive, $\mathsf{Q} \vdash \mathrm{Con}_{\sigma|k}$ and thus $\mathsf{S} \vdash \mathrm{Con}_{\sigma|k}$ for each $k \in \mathbf{N}$. Set $\sigma^* := \sigma(x) \wedge \mathrm{Con}_{\sigma|x}$. Then, σ^* binumerates Q in S. Finally, the claim follows from the arithmetized completeness theorem.[30] $\dashv$

Many of the theories of truth are reflexive; in fact, as we shall see in §4, all the theories of truth we have so far introduced are reflexive. Hence, in many cases, only the truth-free part (indeed Π^0_1 part) of theories does matter to relative interpretability, and relative interpretability is no more fine-grained than the derivability of truth-free theorems.[31]

3.2. Relative truth definability. Now we raise our own proposal, relative truth definability. Informally speaking, that a theory Q is truth-definable in another theory S means that S can translate the truth predicates of Q in such a way that the basic terms are kept unchanged and the concept of truth captured in Q is expressible and its function in Q can be fully simulated by a single formula in S. The formal and precise definition is the following.

DEFINITION 25. Let Q and S be (primitive recursive) theories of truth over languages $\mathcal{L}_{\mathsf{Q}}$ and $\mathcal{L}_{\mathsf{S}}$ respectively, and let $\mathcal{L}_{\mathsf{Q}}$ be $\mathcal{L}_0 \cup \{T_i\}_{i \in I}$ where I is a certain index set; recall that $\mathcal{L}_{\mathsf{Q}}$ may have more than one truth predicates as in the case of $\mathsf{RT}_{<\alpha}$. In order to keep the definition as general as possible, the base theories B of Q and B' of S may be different with the proviso that $\mathcal{L}_{\mathsf{B}} \subset \mathcal{L}_{\mathsf{B}'}$, although we usually consider the cases in which they are the same.

Given a formula θ_i of $\mathcal{L}_{\mathsf{S}}$ for each $i \in I$, we define a translation $\mathcal{T}_{\vec{\theta}}$ from $\mathcal{L}_{\mathsf{Q}}$ to $\mathcal{L}_{\mathsf{S}}$ in the following way:

$$\mathcal{T}_{\vec{\theta}}(\phi) := \begin{cases} \phi & \text{if } \phi \text{ is an atomic formula of } \mathcal{L}_0, \\ \theta_i(x) & \text{if } \phi \text{ is of the form } T_i(x), \\ \neg\mathcal{T}_{\vec{\theta}}(\psi) & \text{if } \phi \text{ is of the form } \neg\psi, \\ \mathcal{T}_{\vec{\theta}}(\psi_0) \vee \mathcal{T}_{\vec{\theta}}(\psi_1) & \text{if } \phi \text{ is of the form } \psi_0 \vee \psi_1, \\ \forall x \mathcal{T}_{\vec{\theta}}(\psi) & \text{if } \phi \text{ is of the form } \forall x\psi. \end{cases}$$

We say Q is *relatively truth definable* (more simply *truth-definable*) in S, or S (relatively) *defines the truth of* Q, when there exists formulae $\theta_i(x)$ of $\mathcal{L}_{\mathsf{S}}$ for each $i \in I$ such that $\mathsf{Q} \vdash \phi \Rightarrow \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi)$ for all $\phi \in \mathcal{L}_{\mathsf{Q}}$. Then, we

[30] See §6 and §8 of [29].

[31] Some non-reflexive theories of truth are known. For example, Cantini's $\mathsf{KF}_{\mathrm{int}}$ and its variant $\mathsf{KF}_{\mathrm{tot}}$ are not reflexive, since they are finitely axiomatizable.

say that $\mathcal{T}_{\vec{\theta}}$ is a *relative truth definition* (or simply *truth-definition*) of Q in S and we call θ_i *truth-defining formula* of the truth predicate T_i of Q in S.

In other words, Q is relatively truth definable in S iff there exists an $\mathcal{L}_0$-conservative relative interpretation of Q in S, or, equivalently, iff Q is a subtheory of a definitional expansion of S. In the literature, this kind of interpretation has been occasionally used for determining proof-theoretic strength of theories and the idea and technique themselves are nothing new; e.g., Feferman gave an $\mathcal{L}_{\mathsf{PA}}$-conservative interpretation of $\mathsf{KF}[\![\mathsf{PA}]\!]$ in Σ^1_1-AC to establish $\mathsf{KF}[\![\mathsf{PA}]\!] \leq \Sigma^1_1$-$\mathsf{AC}$.

We will explain below that relative truth definability is finer-grained than other intertheoretic relations. First of all, we note that relative truth definability is finer-grained than relative interpretability, consistency strength and conservativity of truth-free theorems: a relative truth definition is a relative interpretation; if a theory Q is relatively truth definable in S then the truth-free theorems of the former are included in the latter, i.e., $\mathsf{Q} \subset_{\mathcal{L}_0} \mathsf{S}$, and thus $\mathsf{Q} \leq_{\mathrm{Con}} \mathsf{S}$.

On the other hand, we cannot generally obtain proof-theoretic reducibility from relative truth definability, since the latter has no condition on the theory in which the definability itself is proved, although in practice, at least in all the cases in the present paper, a proof of relative truth definability immediately entails proof-theoretic reducibility.

According to Feferman's original formulation [10], a theory Q is proof-theoretically reducible to S for a set of formulae $\Phi \subset \mathcal{L}_{\mathsf{Q}} \cap \mathcal{L}_{\mathsf{S}}$, if there exists a partial recursive function f such that

(I) for each proof p in Q of a formula $\phi \in \Phi$, $f(p)\downarrow$ and is a proof in S of ϕ,

(II) the formalization of (I) is provable in S.

This definition itself is indeed debatable, since, as was pointed out by Niebergall [33], it forfeits transitivity of proof-theoretic reducibility. However, under a moderate assumption, this problem is resolved. Let us say that Q is proof-theoretically reducible to S for Φ *provably in* a theory T when (I) holds for Q and S and, instead of (II), (I) is provable in T. Then, the following fact due to Niebergall and Rathjen is known:[32]

THEOREM 26. *For any (primitive recursive) extensions* Q *and* S *of* $\mathsf{I}\Sigma_1$, *if* Q *is proof-theoretically reducible to* S *for a primitive recursive set* Φ *provably in* S *and the reduction function* f *can be taken to be primitive recursive, then* Q *is proof-theoretically reducible to* S *for* Φ *provably in* $\mathsf{I}\Sigma_1$.

In practice, the reduction function f has been always primitive recursive in the literature: indeed, Rathjen rather defines proof-theoretic reducibility in [36] with the restriction that f be primitive recursive. In addition, we

[32]The proof of Theorem 26 appeared in [12], which Feferman accredits to Niebergall, whereas Rathjen also pointed out the same fact in [36, §2.5].

usually deal with theories stronger than $\mathsf{I}\Sigma_1$ in the context of foundational discussions and the present paper only considers such theories. Thus we can adopt, as our definition, proof-theoretic reducibility provably in $\mathsf{I}\Sigma_1$ via a primitive recursive reducing function, instead of Feferman's original definition.

Then, if we want relative truth definability to generally imply proof-theoretic reducibility, we need to add some new condition to Definition 25. For example, we can add the following condition $(\flat)$ to Definition 25:

$(\flat)$ θ_i's are primitive recursively specified for each truth predicate T_i of Q, and it is provable in $\mathsf{I}\Sigma_1$ that if $\mathsf{Q} \vdash \phi$ then $\mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi)$; more formally,

$$\mathsf{I}\Sigma_1 \vdash \forall x \in \mathrm{Sent}_{\mathsf{Q}} \left[\mathrm{Prv}_{\mathsf{Q}}(x) \to \mathrm{Prv}_{\mathsf{S}}\left(\mathcal{T}_{\vec{\theta}}(x)\right)\right], \tag{2}$$

where $\mathrm{Prv}_{\mathsf{Q}}$ $(\mathrm{Prv}_{\mathsf{S}})$ are canonical representations of provability in Q (S).

Then, the definition of truth-definability augmented by $(\flat)$ implies proof-theoretic reducibility in general, since we can obtain by a well-known meta-mathematical property of $\mathsf{I}\Sigma_1$ that (2) is equivalent to

$$\mathsf{S} \vdash \forall x \in \mathrm{Sent}_{\mathsf{Q}} \forall y \left[\mathrm{Proof}_{\mathsf{Q}}(y, x) \to \mathrm{Proof}_{\mathsf{S}}\left(g(y), \mathcal{T}_{\vec{\theta}}(x)\right)\right],$$

for some primitive recursive g, where $\mathrm{Proof}_{\mathsf{Q}}(y, x)$ $(\mathrm{Proof}_{\mathsf{S}}(y, x))$ canonically expresses that y is a code of a proof in Q (S) of the formula coded by x. Furthermore, if Q is a first-order classical system without extra inference rules, we only need: e.g.,

$(\flat')$ θ_i's are primitive recursively specified for each T_i of Q, and we can construct a proof of $\mathcal{T}_{\vec{\theta}}(\sigma)$ for each Q-axiom σ primitive recursively on σ.

We think that these additional conditions neither lessen applicability and usability of truth-definability in practice nor harm the philosophical significance as a tool for comparing conceptual aspects of truth theories, although more careful discussions and philosophical evaluations are to be waited and left for another occasion. At any rate, if one does not stick to general implication from relative truth definability to proof-theoretic reducibility, she need not worry too much about the issue argued so far, since relative truth definability immediately implies proof-theoretic reducibility in practice. In fact, in the subsequent proofs of relative truth definability in §4, we first primitive recursively assign a formula θ of $\mathcal{L}_{\mathsf{S}}$ to each truth predicate T of $\mathcal{L}_{\mathsf{Q}}$ and then prove that $\mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\sigma)$ for all Q-axioms σ and $\mathcal{T}_{\vec{\theta}}$ preserves each inference rule of Q. Given such a translation $\mathcal{T}_{\vec{\theta}}$, we can primitive recursively transform any proof of an $\mathcal{L}_0$-sentence in Q to a proof of the same sentence in S.

Whether we add some additional conditions like $(\flat)$ to its definition or not, relative truth definability can be proved to be transitive; note that it is trivially reflexive. Hence relative truth definability induces a pre-ordering among theories of truth; then the classes of mutually truth-definable theories form equivalent classes. Niebergall [33] raised five conditions for a good axiomatization of the reducibility relation ρ: for classical first-order systems Q and S,

PRL1: $\mathsf{Q} \subset \mathsf{S} \Rightarrow \mathsf{Q}\rho\mathsf{S}$;
PRL2: ρ is transitive;
PRL3: if $\mathsf{Q}\rho\mathsf{S}$ then $\mathrm{Con}(\mathsf{S})$ implies $\mathrm{Con}(\mathsf{Q})$;
PRL4: for each finite $\mathsf{Q}' \subset \mathsf{Q}$, if $\mathsf{Q}\rho\mathsf{S}$, there is a finite $\mathsf{S}' \subset \mathsf{S}$ such that $\mathsf{Q}'\rho\mathsf{S}'$;
PRL5: $\mathsf{I\Sigma}_1 \vdash \mathrm{Con}(\mathsf{S}) \rightarrow \mathrm{Con}(\mathsf{Q})$ for finitely axiomatized Q and S with $\mathsf{Q}\rho\mathsf{S}$.

It is easily shown that relative truth definability meets them all.

We would like to argue that relative truth definability is more suitable means than relative interpretability to indicate 'conceptual reducibility' of one theory to another. First, the problems (A)–(D) (in §3.1) of relative interpretability no longer apply to relative truth definability. We have already observed that (C) is not the case for truth-definability; (D) doesn't hold for relative truth definability as we shall see in §4; since the base language and quantification are kept unchanged in relative truth definition, neither (A) nor (B) is the case. Second, some pathological examples like T (i.e., $\mathsf{B} + T(\ulcorner 0 = 0 \urcorner)$) above and UTB are excluded by relative truth definability.

Proposition 27. UTB *is not truth-definable in* T.

Proof. Let T' be T plus $\forall x(x \neq \ulcorner 0 = 0 \urcorner \rightarrow \neg Tx)$. Suppose UTB is truth-definable in T and thus T'. Note that $\mathsf{T}' = \mathsf{B} + \forall x[Tx \leftrightarrow x = \ulcorner 0 = 0 \urcorner]$. Then, if $\mathsf{T}' \vdash \phi$, the formula ϕ' obtained from ϕ by replacing each occurrence of $T(t)$ by $t = \ulcorner 0 = 0 \urcorner$ is derivable in B. This implies that B can define the truth of UTB: this contradicts Tarski's undefinability theorem. ⊣

3.3. Mathematical properties of truth-definability. We have mainly looked at the philosophical significances of relative truth definability, but it is also mathematically and technically useful. In the present subsection, we will see some of its mathematical properties and applicabilities.

First of all, as we have already observed, for theories Q and S, if Q is truth-definable in S, then $\mathsf{Q} \subset_{\mathcal{L}_0} \mathsf{S}$, Q is interpretable in S, $\mathsf{Q} \leq_{\mathrm{Con}} \mathsf{S}$, and $\mathsf{Q} \leq \mathsf{S}$ under some moderate condition like $(\flat)$.

The next proposition immediately follows from the definition.

Proposition 28. (1) *Suppose* Q *is truth-definable in* S *and* $M \models \mathsf{B}$ *can be expanded to a model of* S (*i.e., there exist* $X_i \subset |M|$ *for each truth predicate* T_i *of* S *such that* $(M, \langle X_i \rangle_{i \in I}) \models \mathsf{S}$). *Then* M *is also expandable to a model of* Q.
(2) *Suppose* Q *is truth-definable in* S. *If* Q *is* ω*-inconsistent, so is* S.
(3) *Suppose* Q *is truth-definable in* S. *Then,* Q *plus full induction for* $\mathcal{L}_{\mathsf{Q}}$ *is truth-definable in* S *plus full induction for* $\mathcal{L}_{\mathsf{S}}$.

Thus, relative truth definability preserves useful mathematical properties. Proposition 28 tells us that relative truth definability can distinguish pathological theories like ω-inconsistent theory FS from others, since no ω-consistent theory can define the truth of an ω-inconsistent theory. It also tells that relative truth definability implies 'conservativity in the semantic sense for $\mathcal{L}_0$'; here a theory S is conservative in the semantic sense over Q for $\mathcal{L}_0$ iff the $\mathcal{L}_0$-reduct of every model of Q is expandable to a model of S. McGee [32] argued that conservativity in the semantic sense is to be preferred to and has more philosophical significance than conservativity in the ordinary sense ('conservativity in the proof-theoretic sense' in [32]), since, he claims, there is no metaphysical cost in moving from one theory of truth to another conservative theory in the semantic sense and this move between conservative theories in the semantic sense makes no difference about what the world is like. On the other hand, for instance, $\mathsf{RT}_{<\omega}$ and FS are conservative over each other for $\mathcal{L}_0$ in the ordinary sense but the latter makes a stronger ontological commitment than the former since FS is ω-inconsistent.

In contrast to relative truth definability, relative interpretability and proof-theoretic reducibility satisfy none of (1)–(3) above. It is thus observed that relative truth definability is a finer-grained tool for comparing theories of truth than relative interpretability and proof-theoretical reducibility (and thus consistency strength and conservativity).

We will show one more important feature of relative truth definability in its application to Feferman' *schematic reflective closure* [11]. He introduced this notion as an answer to the question 'which schemata ... in the language of [a theory] ought to be accepted if one has accepted the basic (schematic) axioms and rules of [the theory] ([11, p. 2])'.[33]

DEFINITION 29. Given a base theory B over $\mathcal{L}_0$, let $\mathsf{B}(P)$ denote a theory over $\mathcal{L}_0(P) := \mathcal{L}_0 \cup \{P\}$, where P is a new unary predicate symbol, whose non-logical axioms and rules are just the same as B (i.e., P is a vacuous predicate in $\mathsf{B}(P)$ for which only logical axioms are assumed).

Then, the schematic reflective closure Q^* of a theory Q of truth with a base theory B, is a theory over $\mathcal{L}_{\mathsf{Q}}(P) = \mathcal{L}_{\mathsf{Q}} \cup \{P\}$ whose axioms consist of $\mathsf{Q}[\![\mathsf{B}(P)]\!]$, full induction schema for $\mathcal{L}_{\mathsf{Q}}(P)$, and a new inference rule P-Subst:

$$\frac{\phi(P)}{\phi\big(\hat{x}\psi(x)\big)}\ (P\text{-Subst}), \tag{3}$$

for any formula $\phi(P)$ of $\mathcal{L}_0(P)$ (possibly interspersed with occurrences of P) and a formula ψ of $\mathcal{L}_{\mathsf{Q}}(P)$, where $\phi\big(\hat{x}\psi(x)\big)$ is the result of replacing each

[33]The definition is modified for our current arguments and not exactly the same as Feferman's original one. Feferman's original definition depends on what is regarded as 'schematic axiom'. At any rate, if we focus on arithmetical base theories and only regard induction schema as 'schematic axiom', then Feferman's definition and ours coincide.

occurrence of Pt in ϕ by $\psi[t/x]$ and untangling bound variables.[34] Following Feferman's notation, we sometimes write $\phi(\hat{\psi})$ for $\phi(\hat{x}\psi(x))$.

Notice that $\mathsf{Q}[\![\mathsf{B}(P)]\!]$ contains truth-axioms for the new predicate symbol P. For example, $\mathsf{KF}[\![\mathsf{B}(P)]\!]$ contains the axiom

$$\mathrm{CT}(x) \to \left[P\big(\mathrm{val}(x)\big) \leftrightarrow T(\underset{\cdot}{P}\, x)\right],$$

and each '$\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}$' occurring in **K0**–**K8** is replaced by $\mathrm{Sent}_{\mathcal{L}_\mathrm{T}(P)}$.

Schematic reflective closure can be viewed as an analogue to the Bar rule for sybsystems of second-order arithmetic (cf. [11, §5]). It is not only a philosophically significant framework but also a useful and natural device of extending a given theory. It indeed increases the proof-theoretic strength of theories in many cases (though we will see some exceptional cases in §5).

THEOREM 30 (Feferman [11]). $\mathsf{KF}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$. *Thus* $\mathsf{KF}^*[\![\mathsf{PA}]\!]$ *is also proof-theoretically equivalent to, e.g.*, Σ^1_1-$\mathsf{AC} + \mathrm{BR}$, $\widehat{\mathsf{ID}}_{<\omega}$ *and* ATR_0.[35]

In addition to his first one, Feferman also conjectured that

FEFERMAN'S SECOND CONJECTURE. $\mathsf{DT}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$.

We will later show that this is indeed true.

The next lemma shows that truth-definability carries over to schematic reflective closures.

LEMMA 31. *Let* Q *be a first-order classical system of truth with a base theory* B (*without any extra inference rules such as* NEC *and* CONEC). *Suppose* $\mathsf{Q}[\![\mathsf{B}(P)]\!]$ *is truth definable in* $\mathsf{S}[\![\mathsf{B}(P)]\!]$ *and let* $\vec{\theta}$ *be the truth-defining formulae. Then, the same translation* $\mathcal{T}_{\vec{\theta}}$ *is a truth-definition of* Q^* *in* S^*.

Even for Q *with extra inference rules, if* $\mathcal{T}_{\vec{\theta}}$ *further preserves all the inference rules of* $\mathsf{Q}[\![\mathsf{B}(P)]\!]$ *in the sense that, for each inference rule* $\mathcal{R}$ *of* Q *in the form 'if* $\phi_0, \dots, \phi_n$ *are derived, then infer* $\psi_{\vec{\phi}}$*',* $\mathcal{T}_{\vec{\theta}}$ *satisfies that*

$$\text{if } \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi_0), \dots, \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi_n), \text{ then } \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\psi_{\vec{\phi}}),$$

then Q^* *is truth-definable in* S^* *via* $\mathcal{T}_{\vec{\theta}}$.

PROOF. By induction on the length of derivation. The $\mathcal{T}_{\vec{\theta}}$-translation of each instance of full induction for $\mathcal{L}_\mathsf{Q}(P)$ and each Q^*-axiom trivially holds in S^*. We show that the inference rules are preserved by $\mathcal{T}_{\vec{\theta}}$. Generalization and Modus Ponens are obviously preserved, and so are the other extra rules (if any) by the assumption. For P-Subst, suppose $\mathsf{Q}^* \vdash \phi(P)$. Then, we have $\mathsf{S}^* \vdash \mathcal{T}_\theta\big(\phi(P)\big)$ by IH. Since $\mathcal{T}_\theta$ does not change P, we have $\mathsf{S}^* \vdash \big(\mathcal{T}_{\vec{\theta}}(\phi)\big)(P)$. Thus, for any $\psi \in \mathcal{L}_\mathsf{Q}(P)$, we obtain $\mathsf{S}^* \vdash \big(\mathcal{T}_{\vec{\theta}}(\phi)\big)\big(\widehat{\mathcal{T}_{\vec{\theta}}(\psi)}\big)$ by

[34] As Feferman pointed out, we should not allow arbitrary ϕ of $\mathcal{L}_\mathsf{Q}(P)$ in the premise of P-Subst; otherwise, it may leads to a contradiction. For example, if we allow arbitrary ϕ in KF^*, then we obtain $T(\underset{\cdot}{P}\, x) \leftrightarrow \psi$ for all ψ.

[35] For the definition of $\widehat{\mathsf{ID}}_{<\omega}$, see [9]. For ATR_0, see [38].

P-Subst. Finally, since $\mathcal{T}_{\vec{\theta}}$ preserves the logical connectives and quantifiers, $\mathsf{S}^* \vdash \mathcal{T}_{\vec{\theta}}(\phi(\hat{\psi}))$ is obtained. ⊣

This property is neither possessed by relative interpretability nor proof-theoretic reducibility. $\mathsf{KF}^*[\![\mathsf{PA}]\!]$ is neither proof-theoretically reducible nor relatively interpretable to $(\mathsf{RT}_{<\varepsilon_0})^*[\![\mathsf{PA}]\!]$ (by Lemma 54 below), whereas $\mathsf{KF}[\![\mathsf{PA}(P)]\!]$ is both proof-theoretically reducible to and interpretable (by Theorem 24) in $\mathsf{RT}_{<\varepsilon_0}[\![\mathsf{PA}(P)]\!]$.

Remark. If a theory B meets the conditions (a)–(d) in §2.1 for our base theory, then so does $\mathsf{B}(P)$. Hence, in the present article, whenever we establish the truth definability of Q in S without specifying the base theory, it applies to both base theories B and $\mathsf{B}(P)$ and thus Q^* becomes automatically truth-definable in S^* as well by the last lemma.

3.4. Digression—proof-theoretic ordinal—. The notion of proof-theoretic ordinal figures prominently among measures of the strength of subsystems of second-order arithmetic and various other systems, and the subject called ordinal analysis is now one of the main streams of proof theory (see [36] or [34] for an overview). However, the usual definition of proof-theoretic ordinal cannot be directly applied to theories of truth. Some alternative definitions still seem implausible for theories of truth. Thus the 'proof-theoretic ordinal' cannot be an appropriate measure until some suitable definition is given for theories of truth. We shortly digress to discuss this issue.

Given a subsystem of second-order arithmetic Q, the proof-theoretic ordinal of Q usually denotes its Π^1_1-ordinal: i.e.,

$$\sup\{\alpha \mid \mathsf{Q} \vdash TI(\prec) \wedge otyp(\prec) = \alpha \wedge \prec \text{ is primitive recursive}\},$$

where $TI(\prec)$ expresses the transfinite induction along $\prec$, that is,

$$\forall X\big[\forall x\big(\forall y \prec x(y \in X) \to x \in X\big) \to \forall x \in \mathit{field}(\prec)\big(x \in X\big)\big].$$

This definition presupposes that the language of Q contains second-order parameters. Since they are not contained in the language of truth theories, the above definition should be modified in a certain way.

One such modification for theories without second-order parameters can be found in [36] for example. There Rathjen considers augmenting a theory with a new unary predicate U and redefine

$$TI(\prec) := \forall x\big(\forall y \prec xU(y) \to U(x)\big) \to \forall x \in \mathit{field}(\prec)U(x)\big].$$

However, it remains to be decided which properties are to be assumed for U, and indeed the resulting ordinal depends on which to assume. Suppose that we postulate all the axioms governing truth predicate and mathematical induction even for U. Then, for instance, the resulting proof-theoretic ordinal of $\mathsf{KF}[\![\mathsf{PA}]\!]$ would be $\varphi_{\varepsilon_0}(0)$. On the other hand, when we only postulate induction for U, the proof theoretic ordinal would be ε_0.[36] If we postulate nothing for U then the ordinal would then be ω.

Another possible definition is for it to be the α such that $\mathsf{Q} \equiv \mathrm{PA} + \bigcup_{\beta<\alpha} \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(\beta)$ or $\sup\{\beta \mid \mathrm{PA} + \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(\beta) \subset \mathsf{Q}\}$. However, both are only focusing on the truth-free part, and thus they are unsatisfactory for the same reason as we denied proof-theoretic reducibility or conservativity.[37]

§4. Several results on truth definability. In this section, we will present several results on truth-definability among theories we have introduced in §2.

For simplicity and readability, let $\preceq$ stand for the truth-definability relation: i.e., for theories Q_0 and Q_1, we write $\mathsf{Q}_0 \preceq \mathsf{Q}_1$ if Q_0 is truth-definable in Q_1. Then, $\mathsf{Q}_0 \prec \mathsf{Q}_1$ is for $\mathsf{Q}_0 \preceq \mathsf{Q}_1 \;\&\; \mathsf{Q}_1 \npreceq \mathsf{Q}_0$.

Some results on truth-definability immediately follow from already known or shown facts. First, if Q_0 is a subtheory of Q_1, then the former is trivially truth-definable in the latter (via the identity translation). Hence, for example,

$$\mathsf{TB} \preceq \mathsf{UTB}, \qquad \mathsf{PUTB} \preceq \mathsf{KF}, \qquad \mathsf{KF} \preceq \mathsf{KF}^*.$$

Second, since truth-definability implies conservativity for $\mathcal{L}_0$-sentences, any theory not conservative (for $\mathcal{L}_0$-sentences) over a theory Q is not truth-definable in Q. Hence, for example, the following negative results on truth-definability are already available:

$$\mathsf{TB}, \mathsf{UTB}, \mathsf{TC}^-, \mathsf{KF}^- \nsucceq \mathsf{TC} \nsucceq \mathsf{FS} \nsucceq \mathsf{RT}_{<\varepsilon_0}.$$

[36] I only give a sketch of its proof here. Let us denote the system we are considering by $\mathsf{KF}(U)$. Take a suitable semi-formal system to which $\mathsf{KF}(U)$ is embedded; such a system can be obtained by adding inference rules for the truth predicate T to ω-arithmetic; cf. [2]. In this system, no sentence containing U can be critical in the rules for T. We call a derivation $\mathcal{D}$ in this system *quasi-normal*, when cut is only applied to atomics of the forms Tt and $\neg Tt$. Then, we can show that, if $\prec$ is a well-ordering and $otyp(\prec)$ is limit, $otyp(\prec)$ is always less than or equal to the least length α of quasi-normal derivations of $TI(\prec)$. This is shown in a similar way to Theorem 6.7.2 (Boundedness Theorem) of [34]. It is crucial in its proof that U cannot be critical in any rule for T. For example, in showing the analogue to Lemma 6.6.8 (Boundedness Lemma) of [34] (we can take a fixed interpretation of T in $\mathbf{N}$, say, the least Kripkean fixed-point, since it doesn't depend on the interpretation of U at all), though T may occur in Δ (we are using the same notation of [34]), the positiveness or negativeness of U's occurrences isn't affected by any rule for T and the induction steps work in the same way as the cited lemma; by contrast, in the case of $\mathsf{KF}[\![\mathsf{PA}(U)]\!]$, this argument is no longer valid because U can be critical in the rules for T and the interpretation of T must depend on the interpretation of U. However, we can also show by the standard technique that each derivation $\mathcal{D}$ in $\mathsf{KF}(U)$ can be transformed into a quasi-normal derivation $\mathcal{D}'$ in the semi-formal system of the length $< \varepsilon_0$. This entails that if $\mathsf{KF}(U) \vdash TI(\prec)$ then $otyp(\prec) < \varepsilon_0$.

[37] For another example, Beklemishev [1] introduced Π^0_n-ordinal of theories ($n \in \mathbf{N}$) which can be defined without appealing to second-order parameters. However, this definition is to be denied for the same reason that it only looks at the arithmetical (truth-free) part.

Third, the next proposition follows from Proposition 28, since no theory but FS is ω-inconsistent.[38]

LEMMA 32. *None of the theories introduced so far* (*except* FS *itself*) *can define the truth of* FS.

Fourth, due to Tarski's undefinability theorem, no theory of truth presented so far is truth-definable in PA, since every theory can derive the TB-schema.

In the rest of this section, we will show less trivial results.

LEMMA 33. $\mathsf{KF}^- \not\preceq \mathsf{TC}^-$, *and* $\mathsf{TC}^- \not\preceq \mathsf{KF}^-$. *On the other hand, we have* $\mathsf{TC} \prec \mathsf{KF}$ *since the former is a subtheory of the latter.*

PROOF. The first claim follows from the fact that KF is stronger than TC (cf. Theorem 35 below). For the second claim, let B' be B plus full-induction for $\mathcal{L}_0$. For the sake of contradiction, suppose $\mathsf{TC}^- \preceq \mathsf{KF}^-$. Since a relative truth definition preserves the $\mathcal{L}_0$-part, this entails that $\mathsf{TC}^-[\![\mathsf{B}']\!] \preceq \mathsf{KF}^-[\![\mathsf{B}']\!]$ as well. However, every nonstandard model $M \models \mathsf{B}'$ must be recursively saturated with respect to $\mathcal{L}_0$ (i.e., every recursive type $p(x)$ from $\mathcal{L}_0$ over M is realized in M) whenever it is expandable to a model of $\mathsf{TC}^-[\![\mathsf{B}']\!]$; Lachlan [28] showed this for $\mathsf{B}' = \mathsf{PA}$, but the proof of Lachlan's result (also see [24]) can be directly generalized to an arbitrary base theory B' over $\mathcal{L}_0$ which meets the conditions we posed in §2 and contains full-induction for $\mathcal{L}_0$. On the other hand, since B is arithmetically sound and thus B' has an infinite model, B' has a non-recursively saturated model; we can take the prime model of some suitable complete extension of B', which is not recursively saturated but a nonstandard model of this extension.[39] By Proposition 28 (1), these contradict the fact that any model $M \models \mathsf{B}'$ can be expanded to a model of $\mathsf{KF}^-[\![\mathsf{B}']\!]$ ([2, §5]). ⊣

LEMMA 34. $\mathsf{TB}, \mathsf{UTB} \prec \mathsf{TC}^-$.

PROOF. It is obvious that $\mathsf{TB}, \mathsf{UTB} \preceq \mathsf{TC}^-$, since the former two are subtheories of the latter. We can show by the same argument as the last lemma that the converse fails, since TB and UTB are conservative over B in the semantic sense.[40] ⊣

[38] Recall that we have assumed that B is arithmetically sound and thus has a model over **N**. Then, it is known that all the theories in §2 except FS has a model over **N** and thus ω-consistent.

[39] For the details of the proof, we refer the readers to [24, §8.1, §11.2], where these results are proved for PA but they can be generalized for an arbitrary B with full-induction.

[40] Let B' be B plus full-induction for $\mathcal{L}_0$. We can show that UTB plus full-induction for $\mathcal{L}_{\mathrm{T}}$ is proof-theoretically reducible to B' for $\mathcal{L}_0$ by using partial truth predicates; cf. fn. 5. This gives an alternative proof of this lemma, since $\mathsf{TC}^-[\![\mathsf{B}]\!] \preceq \mathsf{UTB}[\![\mathsf{B}]\!]$ implies $\mathsf{TC}^-[\![\mathsf{B}']\!] \preceq \mathsf{UTB}[\![\mathsf{B}']\!]$ and thus $\mathsf{TC}[\![\mathsf{B}']\!]$ would become truth-definable in $\mathsf{UTB}[\![\mathsf{B}']\!]$ plus full-induction for $\mathcal{L}_{\mathrm{T}}$. In contrast, it is known that UTB plus full induction is not conservative over B' in the semantic sense; only recursively saturated B'-models can be expanded to its model (for the proof, see [24, §15]).

There still remains one open problem on the relative truth definability among conservative theories of truth over B: we do not yet know whether UTB is truth-definable in TB.

THEOREM 35. $\mathsf{RT}_{<\varepsilon_0}$ *is truth-definable in* KF.

We prove a more general fact from which this theorem easily follows. Preliminarily, define an $\mathcal{L}_{\mathrm{T}}$-formula $D^+(x)$ by

$$D^+(x) := \mathrm{Sent}_{\mathcal{L}_{\mathrm{T}}}(x) \land (Tx \lor Fx) \land \neg(Tx \land Fx).$$

Notice that $D^+(x)$ is equivalent to $(Fx \leftrightarrow \neg Tx) \land x \in \mathrm{Sent}_{\mathcal{L}_{\mathrm{T}}}$.

LEMMA 36. *Let* Q *be a theory over* $\mathcal{L}_{\mathrm{T}}$ *which derives the following*:

(i) *The axiom* **K1** (*of* KF);
(ii) $[D^+(x) \land D^+(y)] \to [D^+(\underset{\cdot}{\neg}\, x) \land D^+(x \underset{\cdot}{\lor} y) \land D^+(x \underset{\cdot}{\to} y)]$;
(iii) $[\mathrm{Sent}_{\mathcal{L}_{\mathrm{T}}}(\underset{\cdot}{\forall}\, z.x) \land \forall y D^+ x(y/z)] \to D^+(\underset{\cdot}{\forall}\, z.x)$;
(iv) $D^+(x) \to (T \neg x \leftrightarrow \neg Tx)$;[41]
(v) $D^+(x \underset{\cdot}{\lor} y) \to [Tx \underset{\cdot}{\lor} y \leftrightarrow (Tx \lor Ty)]$;
(vi) $D^+(x \underset{\cdot}{\to} y) \to [Tx \underset{\cdot}{\to} y \leftrightarrow (Tx \to Ty)]$;
(vii) $D^+(\underset{\cdot}{\forall}\, z.x) \to [T \underset{\cdot}{\forall}\, z.x \leftrightarrow \forall y Tx(y/z)]$;
(viii) *The axiom* **K2** (*of* KF).

Then, if $\mathsf{Q} \vdash \mathrm{TI}_{\mathcal{L}_{\mathrm{T}}}(< \alpha)$, *then* $\mathsf{RT}_{<\alpha}$ *is truth definable in* Q.

PROOF. Let h be a binary primitive recursive function such that

$$h(x, \beta) := \begin{cases} x & \text{if } x \in \mathrm{Sent}_\beta \text{ and } \beta < \alpha \\ \ulcorner 0 = 1 \urcorner & \text{otherwise} \end{cases}$$

We write $h_\beta(x)$ for $h(x, \beta)$. Then, by the primitive recursion theorem, we take another primitive recursive function k such that:

$$k(a) := \begin{cases} a & \text{if } a \in \mathrm{AtFml}_{\mathcal{L}_0}, \\ \ulcorner T\big(k \circ h_\gamma(b)\big)\urcorner & \text{if } a = \underset{\cdot}{T}_\gamma\, b \\ & \qquad \text{for } b \in \mathrm{Term}, \\ \underset{\cdot}{\neg} \ulcorner Tk\dot{b}\urcorner & \text{if } a = \underset{\cdot}{\neg}\, b, \\ \ulcorner Tk\dot{b}\urcorner \underset{\cdot}{\lor} \ulcorner Tk\dot{c}\urcorner & \text{if } a = b \underset{\cdot}{\lor} c \\ \qquad (\ulcorner Tk\dot{b}\urcorner \underset{\cdot}{\to} \ulcorner Tk\dot{c}\urcorner, \text{ resp.}) & \qquad (\text{or } a = b \underset{\cdot}{\to} c), \\ \underset{\cdot}{\forall}\, c. \ulcorner Tk(\dot{b}(c/\ulcorner c \urcorner))\urcorner & \text{if } a = \underset{\cdot}{\forall}\, c.b \text{ for } c \in \mathrm{Var}, \\ \ulcorner 0 = 1 \urcorner & \text{otherwise.} \end{cases}$$

We slightly abuse notation here; $\ulcorner T\big(k \circ h_\gamma(b)\big)\urcorner$ and $\underset{\cdot}{\forall}\, z. \ulcorner Tk(\dot{b}(c/\ulcorner c \urcorner))\urcorner$ precisely mean $\ulcorner T\big(k \circ h_\gamma(u)\big)\urcorner(b/\ulcorner u \urcorner)$ and $\underset{\cdot}{\forall}\, c. \ulcorner Tk(\dot{b}(u/\ulcorner c \urcorner))\urcorner(c/\ulcorner u \urcorner)$ where u is a fresh variable; informally, $k(T_\gamma t) = T\big(k \circ h_\gamma(t)\big)$ and $k(\forall z \phi(z)) = \forall z T\big(k \ulcorner \phi(\dot{z})\urcorner\big)$ for a term t and formula ϕ. We notice that if $x \in \mathrm{Sent}_\beta$ for

[41] This (iv) is actually redundant since it follows from the definition of D^+ as was noted.

$\beta < \alpha$ then $kx \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}}$. Then, for each $\beta < \alpha$, we set $\theta_\beta(x)$ to be $Tk(x)$ and we take $\theta_\beta(x)$ as the truth-defining formula of T_β.

We will first show by $\mathrm{TI}_{\mathcal{L}_\mathrm{T}}(\beta)$ for each fixed $\beta < \alpha$ that

$$\mathsf{Q} \vdash (\forall \gamma \leq \beta)\big(\mathrm{Sent}_\gamma(x) \to D^+(kx)\big). \tag{4}$$

Since $\mathsf{Q} \vdash \mathrm{TI}_{\mathcal{L}_\mathrm{T}}(< \alpha)$, it suffices to show that $\mathrm{Sent}_\gamma(x) \to D^+(kx)$ is progressive on γ. Let $a \in \mathrm{Sent}_\gamma$ and suppose the claim up to γ. We show the claim for γ by subinduction on the surface complexity of the sentence coded by a.

The case in which $a \in \mathrm{AtSent}_{\mathcal{L}_0}$ immediately follows from (i) since $k(a) = a$. Suppose $a = \underset{\cdot}{T}_\delta\, x$ for $x \in \mathrm{CT}$ and $\delta < \gamma$. Then, by (viii), it is observed that:

$$\begin{aligned} D^+(ka) &\Leftrightarrow \big[T\ulcorner Tk \circ h_\delta(x)\urcorner \vee F\ulcorner Tk \circ h_\delta(x)\urcorner\big] \\ &\qquad \wedge \big[\neg T\ulcorner Tk \circ h_\delta(x)\urcorner \vee \neg F\ulcorner Tk \circ h_\delta(x)\urcorner\big] \\ &\Leftrightarrow \big[Tk \circ h_\delta(\mathrm{val}(x)) \vee Fk \circ h_\delta(\mathrm{val}(x))\big] \\ &\qquad \wedge \big[\neg Tk \circ h_\delta(\mathrm{val}(x)) \vee \neg Fk \circ h_\delta(\mathrm{val}(x))\big]. \end{aligned}$$

If $\mathrm{val}(x) \in \mathrm{Sent}_\delta$, then $h_\delta(\mathrm{val}(x)) = \mathrm{val}(x)$ and thus $k \circ h_\delta(\mathrm{val}(x)) = k(\mathrm{val}(x))$; the claim follows from IH. Otherwise, $k \circ h_\delta(\mathrm{val}(x)) = \ulcorner 0 = 1\urcorner$; the claim trivially follows from (i). For the cases in which $a = \underset{\cdot}{\neg} b$, $a = b \underset{\cdot}{\vee} c$ or $a = b \underset{\cdot}{\to} c$, we first observe that $D^+(kx) \leftrightarrow D^+(\ulcorner Tk\dot{x}\urcorner)$ for all $x \in \mathrm{Sent}_\gamma$, which is shown by (viii). Then the claim follows from (ii). Next, suppose a is of the form $\underset{\cdot}{\forall} c.b$. We already have $(\forall y)D^+\big(k(b(y/c))\big)$ by SIH, and we also have:

$$D^+\big(\ulcorner Tk(\dot{b}(c/\ulcorner c\urcorner))\urcorner(y/c)\big) \Leftrightarrow D^+\big(\ulcorner Tk(\dot{b}(\dot{y}/\ulcorner c\urcorner))\urcorner\big) \Leftrightarrow D^+k(b(y/c))$$

for each y in Q. This entails the claim $D^+(\underset{\cdot}{\forall} c.\ulcorner Tk(\dot{b}(c/\ulcorner c\urcorner))\urcorner)$ by (iii). Thus, we have shown (4).

Now we can straightforwardly show using (iv)–(viii) that $\mathcal{T}_{\vec{\theta}}$ is indeed a truth definition of $\mathsf{RT}_{<\alpha}$ in Q. We illustrate $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}(\mathbf{R2})$ and $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}(\mathbf{R6})$ for examples. For the former, let $\beta < \alpha$ and $x \in \mathrm{Sent}_\beta$. Then $\mathcal{T}_{\vec{\theta}}(T_\beta \underset{\cdot}{\neg} x)$ is equivalent to $T(\underset{\cdot}{\neg}\ulcorner Tk\dot{x}\urcorner)$. By (4), we already know both $D^+(\underset{\cdot}{\neg}\ulcorner Tk\dot{x}\urcorner)$ and $D^+(kx)$. Hence,

$$T(\underset{\cdot}{\neg}\ulcorner Tk\dot{x}\urcorner) \Leftrightarrow \neg T(\ulcorner Tk\dot{x}\urcorner) \Leftrightarrow \neg Tkx = \mathcal{T}_{\vec{\theta}}(\neg T_\beta x);$$

we thus obtain $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}(\mathbf{R2})$. Next, for the latter, let $\beta < \alpha$ and take any $\delta < \beta$ and $x \in \mathrm{CT}$ with $\mathrm{val}(x) \in \mathrm{Sent}_\delta$. Then, it follows from (viii) that

$$\mathcal{T}_{\vec{\theta}}\big((T_\beta(\underset{\cdot}{T}_\delta\, x)\big) \Leftrightarrow T\big(\ulcorner Tk \circ h_\delta x\urcorner\big) \Leftrightarrow Tk \circ h_\delta(\mathrm{val}(x))$$

Since we have assumed that $\mathrm{val}(x) \in \mathrm{Sent}_\delta$, we have $k \circ h_\delta(\mathrm{val}(x)) = k(x)$ and thus the last formula is equivalent to $\mathcal{T}_{\vec{\theta}}\big(T_\beta \mathrm{val}(x)\big) = Tk(\mathrm{val}(x))$. The other cases can be shown similarly. ⊣

Since KF^- satisfies the conditions (i)–(viii)[42] and $\mathsf{KF} \vdash \mathrm{TI}_{\mathcal{L}_\mathrm{T}}(< \varepsilon_0)$ (by the standard proof), Theorem 35 follows from this lemma. In addition, FKF^- also meets all the conditions (i)–(viii), the next theorem follows.

THEOREM 37. $\mathsf{RT}_{<\varepsilon_0}$ *is truth-definable in* FKF (*and thus* DT).

COROLLARY 38. *If* $\alpha \leq \varepsilon_0$, *then* $\mathsf{DT}^-, \mathsf{KF}^- \not\preceq \mathsf{RT}_{<\alpha}$.[43]

PROOF. We only show the claim for KF^-, but DT^- case is shown in the same manner. When $\alpha < \varepsilon_0$, the claim is obvious since $\mathsf{RT}_{<\varepsilon_0} \preceq \mathsf{KF}$. Let $\alpha = \varepsilon_0$. For the sake of contradiction, suppose KF is truth-definable in $\mathsf{RT}_{<\varepsilon_0}$. Let $\theta(x)$ be the $\mathcal{L}_{\varepsilon_0}$-formula defining the truth of KF. Take a finite $\Gamma \subset \mathsf{RT}_{<\varepsilon_0}$ such that $\Gamma \vdash \mathcal{T}_\theta(\mathbf{K0}) \wedge \cdots \wedge \mathcal{T}_\theta(\mathbf{K8})$. Take the maximum $\beta < \varepsilon_0$ such that T_β occurs in Γ or θ. Then, we have $\Gamma \cup \{\theta\} \subset \mathsf{RT}_{<\beta+1}$ and thus KF would be truth-definable in $\mathsf{RT}_{<\beta+1}$; however, since $\beta + 1 < \varepsilon_0$, it leads to a contradiction. $\dashv$

These techniques can be modified for schematic reflective closures.

LEMMA 39. *Suppose* Q *over* $\mathcal{L}_\mathrm{T}$ *satisfies* (i)–(viii) *of Lemma* 36. *Then,* $\mathsf{RT}_{<\Gamma_0}$ *is truth-definable in* Q^*. *We can take the same truth-definition* $\mathcal{T}_\theta$ *as in the proof of Lemma* 36.

PROOF. Define ordinals β_n for each $n \in \mathbf{N}$ by $\beta_0 = \varepsilon_0$ and $\beta_{n+1} = \varphi_{\beta_n}(0)$. Note that $\Gamma_0 = \lim_{n\to\infty} \beta_n$ and β_n is an epsilon number for all $n \in \mathbf{N}$. By Lemma 36, it suffices to show that $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_\mathrm{T}(P)}(< \beta_n)$ for all $n \in \mathbf{N}$. This is shown by induction on n.

Recall that Q^* contains full induction for $\mathcal{L}_\mathrm{T}(P)$. Thus $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_\mathrm{T}(P)}(< \varepsilon_0)$: we have shown the base step. For the induction step, suppose $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_\mathrm{T}(P)}(< \beta_n)$. Then, by Lemma 36, $\mathsf{RT}_{<\beta_n}[\![\mathsf{B}(P)]\!]$ is truth-definable in Q^*. It follows from Theorem 13 that $\mathsf{RT}_{<\beta_n}[\![\mathsf{B}(P)]\!] \vdash \mathrm{TI}_{\mathcal{L}_0(P)}(< \beta_{n+1})$ and thus $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_0(P)}(< \beta_{n+1})$. In particular $\mathsf{Q}^* \vdash \mathrm{TI}\big(< \beta_{n+1}; P(x)\big)$. Finally, we apply P-Subst and obtain $\mathrm{TI}_{\mathcal{L}_\mathrm{T}(P)}(< \beta_{n+1})$. $\dashv$

THEOREM 40. $\mathsf{RT}_{<\Gamma_0}$ *is truth-definable in* KF^* *and* FKF^* (*and thus in* DT^*).

We now turn to the relative truth definability among type-free theories of truth. Cantini [2] showed that KF can construct a fixed-point of a positive arithmetical operator. Looking closely into his proof, it is observed that only PUTB schema is needed for this construction. Let $\mathcal{L}_0^{(2)}$ be the second-order language which expands $\mathcal{L}_0$ with second-order variables. We have the following.

[42] See Lemma 3.2.4 and 3.2.5 of [11], where $D^+(x)$ is denoted by $D(x)$.

[43] We might expect that ε_0 can be replaced by Γ_0 in this corollary. The author does not yet know whether it is true or false, but we can at least show that there is no truth definition of $\mathsf{KF}^-[\![\mathsf{B}(P)]\!]$ (and $\mathsf{DT}^-[\![\mathsf{B}(P)]\!]$) in $\mathsf{RT}_{<\Gamma_0}[\![\mathsf{B}(P)]\!]$. For there would otherwise be a truth definition KF^* in $(\mathsf{RT}_{<\Gamma_0})^*$ and then we could derive a contradiction in a similar way to the below proof by using Lemmata 39 and 54.

LEMMA 41 (Cantini [2]). *Let $\Phi(X, x, \vec{v})$ be a $\mathcal{L}_0^{(2)}$-formula with no second-order quantifier and with only displayed variables free in which X occurs only positively. Then, there exists a $\mathcal{L}_\mathrm{T}$-formula $\phi(x, \vec{v})$ such that*

$$\mathsf{PUTB}^- \vdash \phi(x, \vec{v}) \leftrightarrow \Phi(\hat{u}\phi(u, \vec{v}), x, \vec{v}),$$

where $\Phi(\hat{u}\phi(u, \vec{v}), x, \vec{v})$ is the result of substituting $\phi(t, \vec{v})$ for each occurrence of $t \in X$ in Φ.

PROOF. By applying the parametrized diagonalization theorem to a $\mathcal{L}_\mathrm{T}$-formula $\Phi(\hat{u}Ta(u/\ulcorner x \urcorner, \vec{v}/\ulcorner \vec{v} \urcorner), x, \vec{v})$ (w.r.t. the variable a), we can find a γ such that

$$\mathsf{PUTB}^- \vdash \gamma(x, \vec{v}) \leftrightarrow \Phi(\hat{u}T\ulcorner \gamma(\dot{u}, \dot{\vec{v}})\urcorner, x, \vec{v}).$$

From the construction of the diagonal formula γ, we can assume that γ is T-positive. Hence, the PUTB schema implies

$$\mathsf{PUTB}^- \vdash \gamma(x, \vec{v}) \leftrightarrow \Phi(\hat{u}\gamma(u, \vec{v}), x, \vec{v}).$$

We can take this γ as the desired ϕ. ⊣

It is observed by Feferman [11] that the truth of KF can be defined by a fixed-point of a certain positive arithmetical operator; Feferman made use of this fact to show that Σ_1^1-AC, within which a fixed-point (but not necessarily the least) of any positive Σ_1^1-operator can be constructed, can define the truth of KF. Since full induction isn't used in Feferman's construction, we indeed carry out the same argument in PUTB^-:

THEOREM 42. KF^- *is truth-definable in* PUTB^-.

Recall that, since a base theory B is arbitrary and thus we can replace B by $\mathsf{B}(P)$, truth-definability carries over to schematic reflective closures (also see the Remark in §3.3).

COROLLARY 43. KF *and* KF^* *are truth-definable in* PUTB *and* PUTB^* *respectively.*

THEOREM 44. $\mathsf{PUTB}^- \equiv \mathsf{B}$, $\mathsf{PUTB} \equiv \mathsf{KF}$ *and* $\mathsf{PUTB}^* \equiv \mathsf{KF}^*$. *In particular, we have* $\mathsf{PUTB}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$.

It follows from Feferman's consistency proof of DT in [13] that the truth of FKF^- is also defined as an arbitrary fixed-point of a certain positive arithmetical operator. Thus, FKF and FKF^- are truth definable in PUTB and PUTB^- respectively. However, as we have noted in §2.5, fixed-points constructible in $\mathsf{PUTB}^{(-)}$ (or Σ_1^1-AC) are not necessarily consistent. Thus, these techniques are not directly applied to obtain that theories with Cons are truth-definable in $\mathsf{PUTB}^{(-)}$. Fortunately, we can nonetheless modify these techniques so that PUTB^- can define the truth of FKF^- + Cons, i.e., DT^-.[44]

THEOREM 45. DT^- *is truth-definable in* KF^- (*and thus in* PUTB^-).

[44] In contrast, it is still open whether PUTB^- can define the truth of KF^- + Cons.

Proof. First, using the recursion theorem, we define (in PA) a preliminary translation I (on the codes of formulae) such that:

$$I(\phi) = \begin{cases} \phi & \text{if } \phi \text{ is } t = s \text{ or } t \neq s, \\ T \underset{\cdot}{I}\, t & \text{if } \phi \text{ is } Tt, \\ T \underset{\cdot}{I} \underset{\cdot}{\neg}\, t & \text{if } \phi \text{ is } \neg Tt, \\ I(\psi) & \text{if } \phi \text{ is } \neg\neg\psi, \\ [I\psi_0 \wedge I\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\neg\psi_1] & \\ \quad \vee [I\neg\psi_0 \wedge I\psi_1 \wedge \neg I\psi_0 \wedge \neg I\neg\psi_1] & \\ \quad \vee [I\psi_0 \wedge I\neg\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\psi_1] & \text{if } \phi \text{ is } \psi_0 \vee \psi_1, \\ I\neg\psi_0 \wedge I\neg\psi_1 \wedge \neg I\psi_0 \wedge \neg I\psi_1 & \text{if } \phi \text{ is } \neg(\psi_0 \vee \psi_1), \\ [I\psi_0 \wedge I\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\neg\psi_1] & \\ \quad \vee [I\neg\psi_0 \wedge \neg I\psi_0] & \text{if } \phi \text{ is } \psi_0 \to \psi_1, \\ I\psi_0 \wedge I\neg\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\psi_1 & \text{if } \phi \text{ is } \neg(\psi_0 \to \psi_1), \\ \forall x(I\psi \wedge \neg I\neg\psi) & \text{if } \phi \text{ is } \forall x\psi, \\ \forall x\big((I\psi \wedge \neg I\neg\psi) \vee (I\neg\psi \wedge \neg I\psi)\big) & \\ \qquad \wedge \exists x I\neg\psi & \text{if } \phi \text{ is } \neg\forall x\psi, \\ 0 & \text{otherwise,} \end{cases}$$

where $\underset{\cdot}{I}$ is the representation of I. Let $\theta(x)$ be

$$TIx \wedge \neg TI \underset{\cdot}{\neg}\, x \wedge FI \underset{\cdot}{\neg}\, x \wedge \neg FIx.$$

Then we will show that this θ defines the truth of $\mathsf{FKF}^- + \text{Cons}$ in KF^-. This can be done in a straightforward but somewhat tedious manner; we will show some typical cases.

First, note that $\mathcal{T}_\theta(Ft)$ is $TI \underset{\cdot}{\neg}\, t \wedge \neg TIt \wedge FIt \wedge \neg FI \underset{\cdot}{\neg}\, t$. Then, it immediately follows that $\mathsf{KF}^- \vdash \mathcal{T}_\theta(\text{Cons}) \wedge \mathcal{T}_\theta(\mathbf{K2})$. By definition of I, we obviously have $\mathsf{KF}^- \vdash \mathcal{T}_\theta(\mathbf{K1}) \wedge \mathcal{T}_\theta(\mathbf{K3})$ as well.

For **FK4**, let $x, y \in \text{Sent}_{\mathcal{L}_T}$. Then, $\mathcal{T}_\theta(Tx \underset{\cdot}{\vee} y)$ is equivalent to

$$\begin{aligned} &\big[(TIx \wedge TIy \wedge FI \underset{\cdot}{\neg}\, x \wedge FI \underset{\cdot}{\neg}\, y) \vee (TI \underset{\cdot}{\neg}\, x \wedge TIy \wedge FIx \wedge FI \underset{\cdot}{\neg}\, y) \\ &\qquad\qquad\qquad \vee (TIx \wedge TI \underset{\cdot}{\neg}\, y \wedge FI \underset{\cdot}{\neg}\, x \wedge FIy)\big] \\ &\wedge \big[\neg TI \underset{\cdot}{\neg}\, x \vee \neg TI \underset{\cdot}{\neg}\, y \vee \neg FIx \vee \neg FIy\big] \\ &\wedge \big[FI \underset{\cdot}{\neg}\, x \vee FI \underset{\cdot}{\neg}\, y \vee TIx \vee TIy\big] \\ &\wedge \big[(\neg FIx \wedge \neg FIy \wedge \neg TI \underset{\cdot}{\neg}\, x \wedge \neg TI \underset{\cdot}{\neg}\, y) \\ &\qquad \vee (\neg FI \underset{\cdot}{\neg}\, x \wedge \neg FIy \wedge \neg TIx \wedge \neg TI \underset{\cdot}{\neg}\, y) \\ &\qquad\qquad\qquad \vee (\neg FIx \wedge \neg FI \underset{\cdot}{\neg}\, y \wedge \neg TI \underset{\cdot}{\neg}\, x \wedge \neg TIy)\big]. \end{aligned}$$

Since the second and third clauses are entailed from the fourth and first respectively, they can be dropped. Then, the resulting formula is equivalent to:

$$\begin{aligned}&\big[TIx \wedge TIy \wedge FI \dot{\neg} x \wedge FI \dot{\neg} y \wedge \neg FIx \wedge \neg FIy \wedge \neg TI \dot{\neg} x \wedge \neg TI \dot{\neg} y\big]\\ &\quad \vee \big[TI \dot{\neg} x \wedge TIy \wedge FIx \wedge FI \dot{\neg} y \wedge \neg FI \dot{\neg} x \wedge \neg FIy \wedge \neg TIx \wedge \neg TI \dot{\neg} y\big]\\ &\quad \vee \big[TIx \wedge TI \dot{\neg} y \wedge FI \dot{\neg} x \wedge FIy \wedge \neg FIx \wedge \neg FI \dot{\neg} y \wedge \neg TI \dot{\neg} x \wedge \neg TIy)\big].\end{aligned}$$

This formula is equivalent to $\mathcal{T}_\theta\big((Tx \wedge Ty) \vee (Tx \wedge Fy) \vee (Fx \wedge Ty)\big)$. We finally obtain $\mathsf{KF}^- \vdash \mathcal{T}_\theta(\mathbf{FK4})$. ⊣

THEOREM 46. DT *and* DT* *are truth-definable in* KF *and* KF* *respectively* (*and thus in* PUTB *and* PUTB* *respectively*).

From the truth-definability of PUTB^- and DT^- in KF^-, we know:

COROLLARY 47. *If* B *is reflexive, then* PUTB^- *and* DT^- *are reflexive.* PUTB *and* DT *are also* (*essentially*) *reflexive since they contain full induction.*

However, it still remains open whether the converses of Theorem 46 hold. Concluding this section, we restate the two open problems for future works.

OPEN PROBLEM 1. Is PUTB^- (or PUTB) truth-definable in DT^- (DT, resp.)?

OPEN PROBLEM 2. Is UTB truth-definable in TB?

§**5. Some applications.** In the present section, we exhibit some applications of the results and techniques on relative truth definability developed so far. First of all, as we have announced, we give positive answers to both the first and the second conjectures of Feferman.

THEOREM 48. $\mathsf{DT}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\varepsilon_0}$ *and* $\mathsf{DT}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$.

PROOF. The upper bound of $\mathsf{DT}[\![\mathsf{PA}]\!]$ immediately follows from Theorem 46. The converse follows from Theorems 12 and 37.

Similarly, the upper bound of $\mathsf{DT}^*[\![\mathsf{PA}]\!]$ is also obtained by Theorem 46, and it follows from Theorem 12 and 40 that this bound is exact. ⊣

For the next application, we consider another kind of iterative compositional theory of truth. We have pointed out that Feferman Logic is a variant of Weak Kleene Logic with a special conditional. Hence, it is natural to think of an iterative compositional theory of truth based on Weak Kleene Logic (with the ordinary conditional definable by '¬' and '∨'). In fact, Feferman himself mentioned such a system in [11]. Although he didn't explicitly provide its axioms, his intended definition should be the following one and we call it WKF.

DEFINITION 49. The theory WKF, an iterative compositional theory of truth with Weak Kleene schema, is defined as FKF minus the special axioms for the conditional: i.e., $\mathsf{FKF} - \mathbf{FK9} - \mathbf{FK10} + \mathbf{K8}$. **K8** must be added instead of **FK9** and **FK10**, since '$A \to B$' is defined simply as '$\neg A \vee B$' in Weak Kleene schema. Similarly, WKF^- is defined as $\mathsf{FKF}^- - \mathbf{FK9} - \mathbf{FK10} + \mathbf{K8}$.

WKF is also important from the viewpoint of the determinate conception of truth. As we noted in fn. 23, DT does not meet strong compositionality. Although there is a good reason for Feferman's restricting strong compositionality on the conditional '$\rightarrow$', it is then natural to consider a determinate theory of truth with full strong compositionality. Such a theory is obtained by replacing the axiom **D11** by the following:

$$\mathrm{Sent}_{\mathcal{L}_\mathrm{T}}(x) \wedge \mathrm{Sent}_{\mathcal{L}_\mathrm{T}} \rightarrow \left[D(x \dot{\rightarrow} y) \leftrightarrow (Dx \wedge Dy)\right].$$

However, we can show in a parallel manner to Theorem 21 that the resulting theory is identical with $\mathsf{WKF} + \mathrm{Cons}$.

In [11, p. 28], Feferman suggested that the proof-theoretic strength of $\mathsf{WKF}[\![\mathsf{PA}]\!]$ is at least as strong as $\mathsf{RA}_{<\varepsilon_0}$. This is indeed correct since WKF meets the conditions (i)–(viii) of Lemma 36 and proves $\mathrm{TI}_{\mathcal{L}_\mathrm{T}}(<\varepsilon_0)$. For the converse, we can show that a simple interpretation which translates '$T(x \dot{\rightarrow} y)$' by '$T(\dot{\neg} x \dot{\vee} y)$' is a truth-definition of WKF^- (+Cons) in FKF^- (+Cons, resp.); note that we use the recursion theorem to define this translation. Hence, by Theorems 15 and 46, we know the bound is exact. Consequently, we have the following:

THEOREM 50. (1) WKF^- *and* $\mathsf{WKF}^- + \mathrm{Cons}$ *are truth-definable in* KF^-.
(2) $\mathsf{WKF}^- \equiv \mathsf{WKF}^- + \mathrm{Cons} \equiv \mathsf{B}$.
(3) $\mathsf{WKF}[\![\mathsf{PA}]\!] \equiv \mathsf{WKF}[\![\mathsf{PA}]\!] + \mathrm{Cons} \equiv \mathsf{RA}_{<\varepsilon_0}$.
(4) $\mathsf{WKF}^*[\![\mathsf{PA}]\!] \equiv \mathsf{WKF}^*[\![\mathsf{PA}]\!] + \mathrm{Cons} \equiv \mathsf{RA}_{<\Gamma_0}$.

We have so far focused on Cons. Then what would happen when we add Comp? As a matter of fact, Comp collapses all the differences among KF, FKF and WKF: that is, we have the following.

LEMMA 51. $\mathsf{KF}^- + \mathrm{Comp}$, $\mathsf{FKF}^- + \mathrm{Comp}$ *and* $\mathsf{WKF}^- + \mathrm{Comp}$ *are all identical.*

PROOF. The proof is straightforward. For example, observe that $Tx \vee Ty$ is equivalent to $(Tx \wedge Fx) \vee (Fx \wedge Ty) \vee (Tx \wedge Ty)$ under Comp. $\dashv$

In the rest of the present section, we will determine the proof-theoretic strength of the schematic reflective closures of $\mathsf{RT}_{<\alpha}$ and FS by using the results and techniques on truth-definability developed so far. As we shall see, schematic reflective closure does not increase proof-theoretic strength in the cases of FS and $\mathsf{RT}_{<\alpha}$ with $\alpha = \omega^\xi$ for some $\xi > 0$.

For each $k \in \mathbf{N}$ and ordinal α, $(\mathsf{RT}_{<\alpha})^* \restriction_k$ denotes the theory obtained from $(\mathsf{RT}_{<\alpha})^*$ by restricting the number of applications of P-Subst in each derivation to at most k times; therefore $(\mathsf{RT}_{<\alpha})^* \restriction_0$ is just identical with $\mathsf{RT}_{<\alpha}[\![\mathsf{B}(P)]\!]$. In the following, we assume that derivations are formulated in Hilbert style in which inference rules are Modus Ponens and P-Subst only.

LEMMA 52. $(\mathsf{RT}_{<\alpha})^* \restriction_{k+1}$ *is proof-theoretically reducible to* $(\mathsf{RT}_{<\alpha+\alpha})^* \restriction_k$ *for* $\mathcal{L}_\alpha(P)$.

PROOF. Take an arbitrary derivation $\mathcal{D} = \langle \theta_0, \dots, \theta_n \rangle$ in $(\mathsf{RT}_{<\alpha})^* \restriction_{k+1}$. Let $\mathcal{D}_0 = \langle \theta_0, \dots, \theta_m \rangle$ $(m \leq n)$ be the initial segment of $\mathcal{D}$ which ends with the first application of P-Subst $\frac{\theta_j(P)}{\theta_j(\hat{\psi})}$ for some $\psi \in \mathcal{L}_\alpha(P)$: thus θ_m is $\theta_j(\hat{\psi})$ $(j < m)$. Then, there exist $\beta, \gamma \leq \alpha$ such that $\theta_0, \dots, \theta_m \in \mathcal{L}_\beta$ and $\psi \in \mathcal{L}_\gamma$. We define a translation $\mathcal{S}^\beta_\psi$ from $\mathcal{L}_\beta(P)$ to $\mathcal{L}_{\gamma+\beta}(P)$, which 'lifts up' the level of truth predicates in each θ_i $(i < m)$ so that ψ can be treated as if it were at the lowest level; note that $\gamma + \beta \leq \alpha + \alpha$. First we define a preliminary translation g^β_ψ (on the codes of formulae) by the recursion theorem such that:

$$g^\beta_\psi(\phi) := \begin{cases} \phi & \text{if } \phi \text{ is an } \mathcal{L}_0\text{-atomic,} \\ \psi(t) & \text{if } \phi \text{ is } P(t), \\ T_{\gamma+\delta}\big(\underset{\cdot}{g}{}^\beta_\psi(t)\big) & \text{if } \phi \text{ is } T_\delta(t), \\ \neg g^\beta_\psi(\theta) & \text{if } \phi \text{ is } \neg\theta, \\ g^\beta_\psi(\theta_0) \vee g^\beta_\psi(\theta_1) & \text{if } \phi \text{ is } \theta_0 \vee \theta_1 \\ \quad \big(g^\beta_\psi(\theta_0) \to g^\beta_\psi(\theta_1), \text{resp.}\big) & \quad (\text{or } \theta_0 \to \theta_1), \\ \forall z. g^\beta_\psi(\theta) & \text{if } \phi \text{ is } \forall z \theta(z), \end{cases}$$

where $\underset{\cdot}{g}{}^\beta_\psi$ is the representation of g^β_ψ. We notice that $\phi \in \mathcal{L}_\delta$ implies $g^\beta_\psi(\phi) \in \mathcal{L}_{\gamma+\delta}$. The translation $\mathcal{S}^\beta_\psi$ is thereby defined as: $\mathcal{S}^\beta_\psi(R\vec{x})$ is $R\vec{x}$ for each atomic formula of $\mathcal{L}_0$, $\mathcal{S}^\beta_\psi(T_\delta x)$ is $T_{\gamma+\delta}\big(g^\beta_\psi(x)\big)$, $\mathcal{S}^\beta_\psi(Px)$ is $\psi(x)$, and $\mathcal{S}^\beta_\psi$ commutes with connectives and quantifiers.

Then, we can straightforwardly show that $\mathsf{RT}_{<\gamma+\beta}[\![\mathsf{B}(P)]\!] \vdash \mathcal{S}^\beta_\psi(\theta_i)$ for all $i < m$. Since $\theta_j(P) \in \mathcal{L}_0(P)$ and $\mathcal{S}^\beta_\psi$ keeps the $\mathcal{L}_0$-part unchanged, we have $\mathcal{S}^\beta_\psi(\theta_j) = \theta_j(\hat{\psi})$ and thus $\mathsf{RT}_{<\gamma+\beta}[\![\mathsf{B}(P)]\!] \vdash \theta_m$. Let $\mathcal{D}_1$ be this derivation of θ_m in $\mathsf{RT}_{<\gamma+\beta}[\![\mathsf{B}(P)]\!]$. Set $\mathcal{D}' := \langle \theta_0, \dots, \theta_{m-1} \rangle * \mathcal{D}_1 * \langle \theta_{m+1}, \dots, \theta_n \rangle$, where $*$ denotes the concatenation of sequences. Then $\mathcal{D}'$ is a derivation of θ_n in $(\mathsf{RT}_{<\alpha+\alpha})^* \restriction_k$. ⊣

LEMMA 53. *Let α be a limit principal ordinal, i.e., $\alpha = \omega^\xi$ for some $\xi > 0$. Then, $(\mathsf{RT}_{<\alpha})^*[\![\mathsf{B}]\!]$ and $\mathsf{RT}_{<\alpha}[\![\mathsf{B}(P)]\!]$ are identical (as theories).*

PROOF. Suppose $(\mathsf{RT}_{<\alpha})^* \vdash \phi$. Since α is limit, there exist $\beta < \alpha$ and $k \in \mathbf{N}$ such that $\phi \in \mathcal{L}_\beta$ and $(\mathsf{RT}_{<\beta})^* \restriction_k \vdash \phi$. Then, it follows from the last lemma that $(\mathsf{RT}_{<\beta\cdot 2^k})^* \restriction_0 \vdash \phi$. Since α is principal, we have $\mathsf{RT}_{<\alpha}[\![\mathsf{B}(P)]\!] \vdash \phi$. ⊣

THEOREM 54. *$(\mathsf{RT}_{<\alpha})^* \restriction_k$ is proof-theoretically reducible to $\mathsf{RT}_{<\alpha\cdot 2^k}$ for $\mathcal{L}_0$-formulae. Hence, in particular, $\mathsf{TC}^* \leq \mathsf{RT}_{<\omega}$ $[\mathcal{L}_0]$.*

PROOF. Suppose $(\mathsf{RT}_{<\alpha})^* \restriction_k \vdash \phi$ for $\mathcal{L}_0$-formula ϕ. By Lemma 52, we have $\mathsf{RT}_{<\alpha\cdot 2^k}[\![\mathsf{B}(P)]\!] \vdash \phi$ by some derivation $\mathcal{D} = \langle \theta_0, \dots, \theta_n \rangle$ (θ_n is equal to ϕ). Let $\beta \leq \alpha \cdot 2^k$ be an ordinal such that $\theta_0, \dots, \theta_n \in \mathcal{L}_\beta$ and let ψ

be an arbitrary $\mathcal{L}_0$-formula (e.g., $(x)_0 = (x)_1$). Take a translation $\mathcal{S}^{\beta}_{\psi}$ from $\mathcal{L}_{\beta}(P)$ to $\mathcal{L}_{\beta}$ in the same way as Lemma 52. Then, we can show that $\mathsf{RT}_{<\beta}[\![\mathsf{B}]\!] \vdash \mathcal{S}^{\beta}_{\psi}(\theta_n)$. Since $\theta_n \in \mathcal{L}_0$ and $\mathcal{S}^{\beta}_{\psi}$ keeps the $\mathcal{L}_0$-part unchanged, we finally have $\mathsf{RT}_{<\alpha\cdot 2^k}[\![\mathsf{B}]\!] \vdash \theta_n$. $\dashv$

THEOREM 55. *Let α be a limit principal ordinal. Then,* $(\mathsf{RT}_{<\alpha})^*$ *is proof-theoretically equivalent to* $\mathsf{RT}_{<\alpha}$ *for $\mathcal{L}_0$-formulae.*

PROOF. Suppose $(\mathsf{RT}_{<\alpha})^* \vdash \phi$ for $\mathcal{L}_0$-formula ϕ. Since α is limit, there exist $\beta < \alpha$ and $k \in \mathbf{N}$ such that $(\mathsf{RT}_{<\beta})^* \restriction_k \vdash \phi$. Then, by the last theorem, we have $\mathsf{RT}_{<\beta\cdot 2^k} \vdash \phi$. Since α is principal, we obtain $\mathsf{RT}_{<\alpha} \vdash \phi$. $\dashv$

In the same way as Corollary 38, we can show that:

COROLLARY 56. *If $\alpha \leq \Gamma_0$, then* $\mathsf{DT}^*, \mathsf{KF}^* \not\geq (\mathsf{RT}_{<\alpha})^*$.

Now we can determine the proof-theoretic strength of FS^*.

THEOREM 57. FS^* *is proof-theoretically equivalent to* $\mathsf{RT}_{<\omega}$. *Thus, in particular,* $\mathsf{FS}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\omega}$ (*and thus* $\equiv \mathsf{ACA}_0 + \mathrm{BR}$; *cf. fn.* 26 *and Theorem* 54).

PROOF. Halbach [17, §5] already showed that $\mathsf{RT}_{<\omega}$ is truth-definable in FS, and thus $(\mathsf{RT}_{<\omega})^*$ is truth-definable in FS^* by Lemma 31.

For the converse, Halbach [17, §5] also gave a truth-definition g_n of FS_n in $\mathsf{RT}_{<2n}$, where FS_n is obtained from FS by restricting the numbers of applications of NEC and CONEC in each derivation to at most $(n-1)$-times respectively. It is observed from his proof that if $(\mathsf{FS}_n)^* \vdash \phi$ then $(\mathsf{RT}_{<2n})^* \vdash g_n(\phi)$; for, it can be shown in the same way as [17] (the extra cases for P-Subst are obvious) that $(\mathsf{FS}_i)^* \vdash \phi$ implies $(\mathsf{RT}_{<2n})^* \vdash g_i(\phi) \wedge \cdots \wedge g_{2n-i}(\phi)$ for each $i \leq n$. This implies that $\mathsf{FS}^* \leq (\mathsf{RT}_{<\omega})^*$ and thus $\mathsf{RT}_{<\omega}$ by Theorem 55.[45] $\dashv$

§6. Conclusion. The results on relative truth definability obtained so far are partially summarized by the following diagram: for $\mathsf{B} = \mathsf{PA}$,[46]

[45]Halbach's truth-definitions were for $\mathsf{B} = \mathsf{PA}$, but they can be straightforwardly generalized for an arbitrary B which meets our conditions for base theories.

[46]As a matter of fact, the diagram applies to an arbitrary base theory B (subject to the conditions (a)–(d) in §2.1). However, in order to establish this, we need to show $\mathsf{KF}^* \not\leq \mathsf{KF}$ and $\mathsf{DT}^* \not\leq \mathsf{DT}$ for arbitrary B in general, but we have not proved them in the present paper. Its proof needs extra technical arguments which go beyond the scope of the present paper. I only give a sketch of the proof here. We first formulate KF (or DT) in a certain suitable semi-formal system with ω-rule. Then, we can show that if KF (DT) $\vdash \phi$ for an arithmetical ϕ then ϕ is provable in the semi-formal system by a derivation with the length $< \varphi_{\varepsilon_0}0$ which contains no application of cut to formulae with T (but may contain cuts to arithmetical formulae). Since KF^* (DT^*) can prove $\mathrm{TI}_{\mathcal{L}_T}(\varphi_{\varepsilon_0}0)$, it can model those infinitary derivations of arithmetical formulae by a formula Tx: i.e., KF^* (DT^*) can prove that if an arithmetical ϕ is provable in KF (DT) then $T\ulcorner\phi\urcorner$ holds. Thus, KF^* (DT^*) can prove the consistency of KF (DT, resp.).

$$
\begin{tikzcd}
\mathsf{PUTB}^* \arrow[r, Leftrightarrow] & \mathsf{KF}^* & \mathsf{DT}^* \arrow[l, squiggly] & \mathsf{RT}_{<\Gamma_0} \arrow[l] & \\
\mathsf{PUTB} \arrow[u] \arrow[r, Leftrightarrow] & \mathsf{KF} \arrow[u] & \mathsf{DT} \arrow[u] \arrow[l, squiggly] & \mathsf{RT}_{<\varepsilon_0} \arrow[u] \arrow[l] & \\
 & & & \mathsf{RT}_{<\omega} \arrow[u] \arrow[r] & \mathsf{FS} \\
 & & & \mathsf{TC} \arrow[u] & \\
\mathsf{PUTB}^- \arrow[uuu] \arrow[r, Leftrightarrow] & \mathsf{KF}^- \arrow[uuu] & \mathsf{DT}^- \arrow[uuu] \arrow[l, squiggly] & \mathsf{TC}^- \arrow[u] & \\
 & \mathsf{PA} \arrow[r] & \mathsf{TB} \arrow[r, squiggly] & \mathsf{UTB} \arrow[u] \arrow[ul] &
\end{tikzcd}
$$

In this diagram, for each theory Q and S, '$\mathsf{Q} \leftarrow \mathsf{S}$' means $\mathsf{Q} \succ \mathsf{S}$, '$\mathsf{Q} \Leftrightarrow \mathsf{S}$' means $\mathsf{Q} \preceq \mathsf{S}$ and $\mathsf{Q} \succeq \mathsf{S}$, and the curly arrows indicate that it is still open whether the converse holds. The theories in the same horizontal line are all proof-theoretically equivalent (for $\mathcal{L}_0$-sentences); but the theories in the 5th and 6th lines are equivalent as well.

Most of the results on relative truth definability so far are one-way and mutual truth definability fails in many cases. This is what we expected, since we introduced the notion in order to distinguish the theories of truth which have the same truth-free theorems but apparently embody different views of truth. Concluding the present paper, we make some comments on the formal results obtained so far.

First, it is observed that ramified truth theories and type-free truth theories are clearly differentiated in view of relative truth definability. Type-free truth theories are not relatively truth definable in ramified truth theories even with much stronger proof-theoretic strength; e.g., $\mathsf{KF}^- \not\preceq \mathsf{RT}_{<\varepsilon_0}$.

Second, relative truth definability can distinguish disquotational theories like UTB and compositional theories like TC^-. However, we have also seen that an iterative compositional theory KF and a disquotational theory PUTB are equivalent in view of relative truth definability. On the one hand, this fact may be regarded as a defect of relative truth definability as a tool for comparing conceptual aspects of theories of truth, since KF and PUTB look different in some respects. For example, although PUTB can define the truth of KF, it can neither prove the compositionality nor iterativity of its own truth predicate in contrast to KF: i.e., PUTB does not prove statements like

$$x, y \in \mathrm{Sent}_{\mathcal{L}_\mathrm{T}} \rightarrow \left[T(x \mathbin{\underset{\cdot}{\vee}} y) \leftrightarrow Tx \vee Ty\right]$$

or

$$x \in \mathrm{CT} \rightarrow \left[T \mathbin{\underset{\cdot}{\neg}} \underset{\cdot}{T}\, x \leftrightarrow T \mathbin{\underset{\cdot}{\neg}} \mathrm{val}(x)\right].$$

For another example, PUTB can be consistently closed by NEC and CONEC, whereas it is not the case for KF; indeed, the resulting theory is even conservative over PUTB for truth-free theorems, and, furthermore, the same holds even when we add uniform versions of NEC and CONEC, in which the premises and conclusions of the rules can contain parameters.[47] On the other hand, their mutual truth definability may still be interpreted to indicate that KF and PUTB are indeed 'conceptually equivalent' despite these 'superficial' differences.

Third, it is suggested that iterative compositional theories of truth with weaker evaluation schema is 'reducible' to those with stronger schema in a stronger sense than mere relative interpretability, proof-theoretic reducibility and others. But it is still open whether this reduction is one-way. The solution to the Open Problem 1 would answer the question whether iterative compositional theories of truth with different schemata indeed give different or incompatible conceptions of truth.

Fourth, as we have seen in §5, relative truth definability is useful in its applications. In particular, relative truth definability makes it easier to compare the schematic reflective closures of theories and mutual truth definability gives more information on model-theoretic features of theories of truth, since it entails conservativity in the semantic sense.

Acknowledgements. This research was supported by a Domus A Scholarship jointly awarded by Merton College and Clarendon Fund. I am heavily indebted to Volker Halbach for his many useful comments and suggestions on earlier drafts of this paper. Indeed the idea of making use of relative truth definability for comparing theories of truth came from discussions with him. I also owe many thanks to Solomon Feferman for his helpful comments and hints and for sending his early draft of [13] which led me to the present paper. I would like to thank the referee for fruitful suggestions and Graham Leigh and Kentaro Sato for their useful comments as well.

REFERENCES

[1] L. D. Beklemishev, *Proof-theoretic analysis by iterated reflection*, ***Archive for Mathematical Logic***, vol. 42 (2003), pp. 515–552.

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

[47] The proofs of these facts on PUTB can be found in [21].

[3] ———, *A theory of formal truth arithmetically equivalent to ID*1, ***The Journal of Symbolic Logic***, vol. 55 (1990), pp. 244–259.

[4] C. Cieśliński, *Deflationism, conservativeness and maximality*, ***Journal of Philosophical Logic***, vol. 36 (2007), pp. 695–705.

[5] D. Davidson, *The folly of trying to define truth*, ***Journal of Philosophy***, vol. 93 (1996), pp. 263–278, reprinted in (S. Blackburn and K. Simmons, editors) ***Truth***, Oxford University Press, Oxford, 1999, pp. 308–322.

[6] S. Feferman, *Reflecting on incompleteness*, handwritten notes, March 1987, 39 pp.

[7] ———, *Systems of predicative analysis*, ***The Journal of Symbolic Logic***, vol. 29 (1964), pp. 1–30.

[8] ———, *Systems of predicative analysis, II: Representations of ordinals*, ***The Journal of Symbolic Logic***, vol. 33 (1968), pp. 193–220.

[9] ———, *Iterated inductive fixed-point theories: Application to Hancock's conjecture*, ***Patras logic symposion*** (G. Metakides, editor), North-Holland, Amsterdam, 1982, pp. 171–196.

[10] ———, *Hilbert's program relativized: Proof-theoretical and foundational reductions*, ***The Journal of Symbolic Logic***, vol. 53 (1988), pp. 364–384.

[11] ———, *Reflecting on incompleteness*, ***The Journal of Symbolic Logic***, vol. 56 (1991), pp. 1–49.

[12] ———, *Does reductive proof theory have a viable rationale?*, ***Erkenntnis***, vol. 53 (2000), pp. 63–96.

[13] ———, *Axioms for determinateness and truth*, ***The Review of Symbolic Logic***, vol. 1 (2008), pp. 204–217.

[14] H. Friedman and M. Sheard, *An axiomatic approach toward self-referential truth*, ***Annals of Pure and Applied Logic***, vol. 33 (1987), pp. 1–21.

[15] P. Hájek and P. Pudlak, ***Metamathematics of first-order arithmetic***, Springer, Berlin, 1993.

[16] V. Halbach, *Axiomatic theories of truth*, ***Stanford encyclopedia of philosophy***, URL: `http://plato.stanford.edu/entries/truth-axiomatic`.

[17] ———, *A system of complete and consistent truth*, ***Notre Dame Journal of Formal Logic***, vol. 35 (1994), pp. 311–327.

[18] ———, ***Axiomatische Wahrheitstheorien***, Akademie Verlag, Berlin, 1996.

[19] ———, *Conservative theories of classical truth*, ***Studia Logica***, vol. 62 (1999), pp. 353–370.

[20] ———, *Truth and reduction*, ***Erkenntnis***, vol. 53 (2000), pp. 97–126.

[21] ———, *Reducing compositional to disquotational truth*, ***The Review of Symbolic Logic***, vol. 2 (2009), pp. 786–798.

[22] ———, ***Axiomatic theories of truth***, Cambridge University Press, Cambridge, forthcoming.

[23] V. Halbach and L. Horsten, *Axiomatizing Kripke's theory of truth*, ***The Journal of Symbolic Logic***, vol. 71 (2006), pp. 677–712.

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

[25] J. Ketland, *Deflationism and Tarski's paradise*, ***Mind***, vol. 108 (1999), pp. 69–94.

[26] H. Kotlarski, S. Krajewski, and A. H. Lachlan, *Construction of satisfaction classes for nonstandard models*, ***Canadian Mathematical Bulletin***, vol. 24 (1981), pp. 283–293.

[27] S. Kripke, *Outline of a theory of truth*, ***Journal of Philosophy***, vol. 72 (1975), pp. 690–716.

[28] A. Lachlan, *Full satisfaction classes and recursive saturation*, ***Canadian Mathematical Bulletin***, vol. 24 (1981), pp. 295–297.

[29] P. Lindström, ***Aspects of incompleteness***, 2nd ed., Lecture Notes in Logic, 10, A K Peters, Ltd, Massachusetts, 2003.

[30] V. McGee, *How truthlike can a predicate be? A negative result*, ***Journal of Philosophical Logic***, vol. 14 (1985), pp. 399–410.

[31] ———, *Maximal consistent sets of instances of Tarski's schema* (T), ***Journal of Philosophical Logic***, vol. 21 (1992), pp. 235–241.

[32] ———, *In praise of the free lunch: Why disquotationalists should embrace compositional semantics*, ***Self-reference*** (T. Bolander et al., editors), CSLI Publications, Stanford, 2006, pp. 95–120.

[33] K. Niebergall, *On the logic of reducibility: Axioms and examples*, ***Erkenntnis***, vol. 53 (2000), pp. 27–61.

[34] W. Pohlers, ***Proof theory: The first step into impredicativity***, Springer, Berlin, 2009.

[35] M. Rathjen, *The role of parameters in bar rule and bar induction*, ***The Journal of Symbolic Logic***, vol. 56 (1991), pp. 715–730.

[36] ———, *The realm of ordinal analysis*, ***Sets and proofs*** (S. Cooper and J. Truss, editors), Cambridge University Press, Cambridge, 1999, pp. 219–279.

[37] M. Sheard, *A guide to truth predicates in the modern era*, ***The Journal of Symbolic Logic***, vol. 59 (1994), pp. 1032–1054.

[38] S. Simpson, ***Subsystems of second order arithmetic***, Springer, Berlin, 1999.

[39] G. Takeuti, ***Proof theory***, 2nd ed., North-Holland, Amsterdam, 1987.

[40] A. Tarski, *The semantic conception of truth and the foundation of semantics*, ***Philosophy and Phenomenological Research***, vol. 4 (1944), pp. 341–376, reprinted in (A. Tarski, editor) ***Alfred Tarski, Collected Paper 2***, Birkhauser, Basel, 1986, pp. 665–699.

[41] ———, *The concept of truth in formalized languages*, ***Logic, semantics, metamathematics*** (A. Tarski, editor), Hackett, Indianapolis, 2nd ed., 1983, pp. 152–278.

MERTON COLLEGE
OXFORD OX1 4JD, UK
E-mail: kentaro.fujimoto@philosophy.ox.ac.uk