THE REVIEW OF SYMBOLIC LOGIC
Volume 10, Number 2, June 2017

EQUIVALENCES FOR TRUTH PREDICATES

CARLO NICOLAI
Munich Center for Mathematical Philosophy, LMU Munich

Abstract. One way to study and understand the notion of truth is to examine principles that we are willing to associate with truth, often because they conform to a pre-theoretical or to a semi-formal characterization of this concept. In comparing different collections of such principles, one requires formally precise notions of inter-theoretic reduction that are also adequate to compare these conceptual aspects. In this work I study possible ways to make precise the relation of conceptual equivalence between notions of truth associated with collections of principles of truth. In doing so, I will consider refinements and strengthenings of the notion of relative truth-definability proposed by Fujimoto (2010): in particular I employ suitable variants of notions of equivalence of theories considered in Visser (2006) and Friedman & Visser (2014) to show that there are better candidates than mutual truth-definability for the role of sufficient condition for conceptual equivalence between the semantic notions associated with the theories. In the concluding part of the paper, I extend the techniques introduced in the first and show that there is a precise sense in which ramified truth (either disquotational or compositional) does not correspond to iterations of comprehension.

§1. The project. Comparing theories is one of the fundamental activities of the scientist and of the philosopher. Often this comparison is carried out via formalization, and there is a great deal of controversy on how to properly formalize scientific or mathematical theories.[1] In this work I focus on a narrower target and investigate notions of reduction and equivalence involving deductive systems obtained by extending a base arithmetical theory with principles of truth. This description is deliberately vague at this stage: I will in fact relate either two theories of truth or a theory of truth and an extension of the base theory with suitable comprehension schemata. In the rest of this introduction, I will motivate the project in its twofold nature.

1.1. Truth and conceptual equivalence. Recent works on truth and semantic paradoxes fall in two broad categories. On one side we find works that aim at describing a class of models for languages endowed with a truth predicate. These construction are carried out by defining the notion of truth in a more powerful metatheory—usually a system of analysis in the case of an arithmetical base theory. We call this approach the *semantic* approach (for an overview, see Field (2008)). On the other side authors that give priority to rules and principles characterizing a primitive notion of truth. On this approach one focuses on formal systems that generate truths (and falsities). The latter approach is widely known as the *axiomatic* approach (Halbach, 2014; Horsten, 2011).

Received: May 22, 2016.
2010 *Mathematics Subject Classification*. 03F25, 03F30, 03F35, 03A05.
Key words and phrases. Formal Theories of Truth, interpretability, sameness of theories, arithmetical comprehension.

[1] See Lutz (2016) for an up to date account of the debate in the context of philosophy of science.

© Association for Symbolic Logic, 2017

doi:10.1017/S1755020316000435

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

That there can be fruitful interactions between the two modus operandi is clear since Tarski's seminal work.[2] Collections of principles and rules usually form succinct and accessible bases for motivating, evaluating, criticizing some approaches to semantic paradoxes. Semantic constructions are often indispensable in the creative act of articulating one such approach.[3]

In order to characterize the notion of truth arising from a semantic construction, one usually fixes a specific model—e.g., the minimal fixed point of Kripke's theory in Strong Kleene logic from Kripke (1975)—and reads off the properties of the truth predicate in this model. If one, by contrast, gives priority to the analysis of systems generating truths, the vicinity to a semantic construction may be part of the characterization of the system's truth predicate, but also other features of this truth predicate need to be part of the analysis. For instance, the vicinity of a system of truth T to a class of models Σ may be evaluated, as proposed by Fischer, Halbach, Kriener, & Stern (2015), by considering ω-models. On this approach, T will capture Σ if and only if[4]

$$(\mathbb{N}, S) \vDash T \Leftrightarrow (\mathbb{N}, S) \text{ belongs to } \Sigma, \tag{1}$$

where S is the extension of the truth predicate.

However, given some S, Σ, there are several examples of systems T_1, T_2 satisfying (1) that cannot be plausibly associated to the same notion of truth.[5] This suggests that, in order to characterize the notion of truth associated with a system of truth, we need to integrate the vicinity to a semantic construction—that may still not necessarily be identified with the criterion of sameness of ω-models of (Fischer et al., 2015)—with further, complementary criteria. A list of such criteria for systems of truth will be provided in §2.

Before sketching the project any further, a note on terminology. We will employ interchangeably the expressions 'notion of truth', 'concept of truth, and 'conception of truth' in relation to the outcome of the reflection on the properties of a truth predicate as

---

[2] Cf. Fischer et al. (2015) and Halbach (2014) for an in-depth treatment of the relationships between the two approaches.

[3] A Kantian aphorism by Hannes Leitgeb may help in depicting the situation: it is often the case that axiomatic truth without semantic truth is empty, whereas semantic truth without axiomatic truth is blind.

[4] For simplicity, we consider theories in classical logic, so $\vDash$ is the classical satisfaction relation.

[5] For instance, let us consider the theories extending a base syntax theory B—that can be taken to be a weak arithmetical system such as *EA* considered below—with typed Tarski biconditionals of the form

$$\text{`}\varphi\text{' is true if and only if } \varphi$$

with φ a sentence of $\mathcal{L}_B$, and the theory obtained by extending B with full, compositional axioms of the form

> for all sentences φ, ψ of $\mathcal{L}_B$, the sentence 'φ and ψ' is true if and only if 'φ' is true and 'ψ' is true.

The two theories are different in many respects: the latter theory is capable of proving many more general claims than the first, it is compositional and not disquotational as the former, and its notion of truth is not reducible to the logical resources of B. Moreover, the compositional theory will *not* in general be relatively interpretable in B, for finitely axiomatizable B, and not conservative over it if the right amount of inductive reasoning with truth is available. Nonetheless, the two truth theories have the same models based on $\mathbb{N}$—that is a pair $(\mathbb{N}, S)$ where S is the unique set of truths for $\mathcal{L}$—so they capture the same semantic construction according to the criterion above and, if the 'reductionist' picture is endorsed, their notions of truth coincide.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

depicted by a theory of truth. In all cases we *intend a cluster of conditions satisfied by the truth predicates of the languages of our theories*. In this sense a notion (conception, concept) of truth is primarily not a psychological notion, and so my focus diverges from the classical ontological and scientific questions on concepts that may be found in the literature.[6]

At the informal level, we want to address the question: when the concepts of truth can be considered equivalent? Still intuitively, our guiding answer is: when they equally possess or lack a number of crucial properties. This characterization is of course still unsatisfactory. In the first place we need the additional assumption that conceptions of truth are given by formal theories, and formal systems in particular. Moreover, the relevant properties that two theories of truth may or may not possess are dependent on one's stance on truth: an instrumentalist may regard provability of truth-free consequences as the only criterion for equivalence of two truth predicates, whereas someone who is interested in the purely truth theoretic principles of the theories may completely disregard their consistency strength. In addition, one would like to integrate the mutual preservation of some features of theories of truth qua mathematical objects into a more general framework of equivalence of formal systems. It would be at least puzzling to realize that systems T_1 and T_2 are formally equivalent in a strong sense but their truth predicates fare differently with respect to some of the characterizing, quasi-formal desiderata generally imposed on notions of truth.

In the first part of this work I will therefore investigate the possibility of finding a formal counterpart to the informal relation of 'conceptual equivalence' just sketched. This is not an easy nor an original task: some authors, notably (Fujimoto, 2010) and (Halbach, 2014), have tried to come up with notions of reductions up to the task. In particular, by applying the notion of *relative truth-definability*—that is a relative interpretation that keeps the syntactic vocabulary unchanged—and variations thereof, they aimed at capturing the notion of *conceptual equivalence* of truth predicates (see §1 of Fujimoto (2010)), or comparing the *conceptual strength* of two (or more) truth predicates. In the first part of this paper I will show that the notions of (proof-theoretic reductions) that are commonly employed in the literature, including relative truth-definability, do not suffice to adequately characterize the relation of 'conceptual equivalence of truth predicates'. In the second part I will consider stricter notions of reduction that may represent a sufficient condition for two theories' truth predicates to be conceptually equivalent, although arguably not a necessary one due to their strictness.

1.2. Classes and truths. In formal terms, the result of extending a base arithmetical theory with truth axioms has been often regarded as another, perhaps more succint way of extending it with further ontological assumptions on the existence of sets of natural numbers. Famously (Feferman, 1991) has employed truth axioms as a device to investigate the limits of predicativity given the natural numbers. This programme finds its roots in the inter-reducibility of suitable truth axioms with certain fragments of ramified analysis. These mutual reductions may be of interest for multiple reasons. In general, relating truth with certain forms of membership may help in harmonizing analyses of semantic, set-theoretic, and property-theoretic paradoxes. Moreover, trading-off ontological commitment to sets with semantic commitment to truth axioms may be attractive for philosophers interested

[6] For various approaches to the notion of concept in philosophy and cognitive science, we refer to (Laurence & Margolis, 1999). Woodfield (1991) also considers conceptions, but his notion is very close to the usual notion of concept employed in psychology.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

in reducing assumptions on the existence of sets to syntactic objects and ideology (Halbach, 2014).[7]

The strict notions of reduction that will be investigated in the first part of the paper will shed light on the folklore reductions between suitable truth axioms and certain forms of comprehension—the ones considered by Halbach and Feferman, for instance—and yield a surprising picture in which semantic and set-theoretic assumptions *fail to be equivalent*. In particular, in this picture ontological commitments seem to be reducible to semantic assumptions but not vice versa.

§2. Truth predicates and theoretical equivalence. In this section I extract from the recent literature some desiderata for adequacy of systems of truth (I mainly refer to Leitgeb (2007), Halbach & Horsten (2015)) to find plausible candidates for conceptual properties of a theory's truth predicate besides the vicinity to some semantic construction. The criteria seem to oscillate between genuinely truth theoretic, or 'conceptual' ones, and criteria on the systems *qua* mathematical objects, although in the axiomatic context the boundaries between the two kinds are often very difficult to trace. In the following list, T and W will denote theories extending a base theory B with one or more truth predicates. The fundamental idea behind the listing is that it cannot be the case T and W are seen as conceptually equivalent and yet they disagree on some of the desiderata below. It should be clear that the list is by no means exhaustive: I simply appeal to criteria that acquired some general consensus to guide our formal analysis.

Ontological commitments. This requirement can be paraphrased as follows: suppose T and W display the same conception of truth. If T enables one to interpret the objects of the domain to which truth is ascribed as the intended bearers of truth, typically sentence types, also W should do so. A more general formulation of this criterion may be as follows: sameness of conception entails preservation of the ontological commitments of the theories.[8] The very possibility of comparing truth predicates, let alone equating them, appears to be rooted in the possibility of applying them to structurally similar linguistic-mathematical objects. It is known, for instance, that there are theories of truth that do not have ω-models, such as FS (cf. Halbach (2014); Friedman & Sheard (1987)). It would be embarrassing, for any satisfactory notion of equivalence of conception of truth, to pair FS with theories that admit standard models.

Truth-theoretic generalizations. The provability of generalizations involving the truth predicate may be seen as a formal rendering of the requirement, for the truth predicate, of adequately 'expressing infinite conjunctions' or allowing for 'semantic ascent'.[9] In other

[7] For instance, in Halbach (2014), we read:

> ...I expect that not only truth theories can be applied to eliminate ontological commitments, but that the work on proof theory also sheds light on how semantic and ontological commitment are related. [...] In particular, the reductions of second-order theories to truth theories may be taken as evidence that ontological commitment can be replaced with ideological or semantic commitment, or perhaps even that there is no very clear distinction between the two kinds of commitment. (Halbach, 2014, p. 318)

[8] Preservation of ontological commitment may also be intended not as *identity* of the syntactic universe, but as isomorphism of the structures of the 'objects of truth'. I will allow for this possibility, although I will not focus on similar cases.

[9] Some authors think that equating 'expressing infinite conjunctions' with 'proving universally quantified statements involving truth' is a hasty move. This is the case of (Halbach, 1999), for instance. Some others, such as (Cieśliński, 2010), find the equation harmless.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

words if the notions of truth associated with T and W are conceptually equivalent, we expect the existence of an effective procedure that enables one to translate (provable) generalizations involving the truth predicate(s) of T, e.g., stating that all members of a countable set $\mathcal{S}$ of formulas are true, into general claims involving *the truth predicate(s)* of W, in which the formulas expressing membership in $\mathcal{S}$ in W and T apply to the same syntactic objects. The requirement in italic should not be underestimated: we will consider below cases of theories that satisfy the condition just stated only by means of resources other than truth.

Reducibility. The debate over truth-theoretic deflationism has prompted an extensive literature on the reducibility of semantic resources to the underlying object theory—for an overview, cf. Horsten (2011). There are at least three notions of reductions that may be considered. The first is definability: it is obvious that this is not available in the case of truth. The second is conservativity: a notion of truth may help in reaching consequences that were not provable by the object theory only. The third is relative interpretability: although the notion of truth is not definable in the syntactic base theory, its behaviour may be replicated by notions that work on a suitable relativization of the syntactic domain. Conservativity and interpretability, besides having deep philosphical implications, are measures of strength of the theories of truth. We require them to be preserved in adequate formal renderings of conceptual equivalence of truth predicates.

Compositionality. In general, we welcome the possibility of understanding the truth of a compound sentence only by knowing the truth values of the compounding sentences (for as many objects to which the truth predicate can be applied as we can find). For the truth predicates of T and W to display the same conception of truth, both theories should have this property: if $Tr_T(\varphi \wedge \psi)$ can be understood—in T—as '$Tr_T{\ulcorner}\varphi{\urcorner}$ and $Tr_T{\ulcorner}\psi{\urcorner}$, also $Tr_W({\ulcorner}\varphi \wedge \psi{\urcorner})^*$ (where $(\cdot)^*$ is an appropriate isomorphism between the syntactic universes), should be understood analogously in W. It cannot be the case, for instance, that T is able to decompose sentences compositionally only for an initial portion of the syntactic universe (perhaps only for standard syntactic objects), whereas W can do so for *any* sentence in the domain of our quantifiers. As a consequence, this criterion enables us to separate between disquotational and compositional theories.

Symmetry. Under the assumption that T features one single truth predicate Tr, the internal theory of T is defined as the set of φ such that $T \vdash Tr{\ulcorner}\varphi{\urcorner}$. The external theory of T simply the set of φ such that $T \vdash \varphi$. When the external and the internal theories of T coincide, T is said to be symmetric. If T and W embody equivalent conceptions of truth and if T can be consistently made symmetric, this should also be possible for W. A mismatch between the internal and the external theories, although often unavoidable, is an unpleasant sign: it may suggest that the meaning externally assigned to logical connectives shifts once we move under the scope of the truth predicate.

Finite axiomatizability (*with no additional resources*). Often accompanied by the compositionality requirement, the finite axiomatizability of a theory of truth is sometimes related to the learnability, by human beings, of truth values for infinitely many sentences starting with finite instructions.[10] It would appear at least questionable for T and W being considered conceptually equivalent, but T be presented as a finite set of clauses whereas W as an infinite list of instructions. This desideratum is often stressed in relation to Davidson's program, but it is not clear why the requirement cannot be extended to *recursive*, instead of finite, sets of formulas. After all, isn't the schema $\varphi \rightarrow \varphi$ learnable, even though it has infinitely many instantiations? Perhaps a distinction can be made by instances of schemata characterizing logical vocabulary only and instances of schemata concerning nonlogical

[10] This is a famous Davidsonian theme. See Davidson (1984).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

vocabulary, but it's not our intention here to follow this line of reasoning any further. Since we are dealing with formal systems of truth, in fact, there is enough evidence supporting the claim that the truth predicate of a finitely axiomatizable theory may considerably differ from the truth predicate of a non finitely axiomatizable theory; for instance, the former may not be reducible to the object theory, whereas the latter may well be.[11]

Type restrictions. We will require the (non) applicability of the truth predicate to sentences containing semantic vocabulary to be part of the conception of truth arising from some axiomatic theory. That said, it is not immediately clear what the distinction between typed and type-free truth may be.[12] We shall be content with the following condition: T will be dubbed as type-free if it proves the truth of a *particular* sentence containing the truth predicate.

2.1. Theories, base theories, theories of truth. Theories of truth require a theory of the objects to which truth applies: in our case eternal, context-free sentence-types. In turn it is well-known that, modulo synonymy (cf. §2.2), direct axiomatizations of syntax or finite sets coincide with suitable arithmetical theories. The main role of the underlying syntax theory is to state the relevant syntactic notions and operations needed for truth ascriptions and to prove their relevant properties: to this purpose, the standard choice of PA as base theory seems unnecessarily strong. In this work I will employ elementary arithmetic EA as base theory: it is a properly weaker theory than PA or $\mathsf{I}\Sigma_1$.[13] EA captures in fact in a direct and natural way the standard development of the syntax of first-order theories, as carried out for instance in Feferman (1960); Smorynski (1977) for PA and PRA.[14] More on the choice of EA can be found in §4.1.

Officially, EA is formulated in a first-order, relational language $\mathcal{L}$ with logical constants in $\{\neg, \wedge, \forall\}$ featuring finitely many relation symbols $\mathsf{Z}(x)$ ('x is identical to zero'), $\mathsf{S}(x, y)$ ('y is the successor of x'), $\mathsf{E}(x, y)$ ('y is 2^x'), $\mathsf{A}(x, y, z)$ ('z is the sum of x and y'), $\mathsf{M}(x, y, z)$ ('z is the product of x and y'). With $x \neq y$, expressions of the form $(\forall x \leq y)(\varphi(x))$ and $(\exists x \leq y)(\varphi(x))$ are said to be obtained from $\varphi(v)$ by *bounded quantification*. The class of formulas of $\mathcal{L}$ that contain only bounded quantifiers will be referred to as the class of *elementary* formulas. The set of axioms of EA contains, besides the logical axioms of predicate logic with identity formulated in a Hilbert-style calculus, the functionality (including totality) axioms for the relation symbols just introduced and axioms corresponding to their recursive clauses such as (the universal closures of)

$$\mathsf{Z}(y) \wedge \mathsf{S}(y, z) \rightarrow (\mathsf{E}(x, z) \leftrightarrow x = y)$$
$$\mathsf{S}(x_0, x_1) \wedge \mathsf{A}(z_0, z_0, z_1) \rightarrow \big(\mathsf{E}(x_0, z_0) \leftrightarrow \mathsf{E}(x_1, z_1)\big).$$

[11] We discuss the finite axiomatizability criterion further in §4.1.

[12] Halbach (2014, chap. 10) proposes the satisfaction of the criteria as sufficient conditions for a system of truth T to be typed:

1. Only sentences not containing used or mentioned occurrences of the truth predicate should be deemed true by the theory;
2. the theory should not rule out the possibility of picking any set definable in the base theory as extension of the truth predicate restricted to sentences containing the truth predicate itself.

[13] More precisely, the consistency of EA can be proved already in PRA, which is known to be conservatively extended by $\mathsf{I}\Sigma_1$ for Π_2-formulas.

[14] Yet, we are far from being close to the theoretical lower bounds. For this purpose, one might choose a theory interpretable in Q as in Nicolai (2016a).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

In addition, we have the schema of bounded induction for elementary (Δ_0) formulas:

$$\mathsf{Z}(x) \wedge \varphi(x) \wedge \forall y, y_0\, (\varphi(y) \wedge \mathsf{S}(y, y_0) \rightarrow \varphi(y_0)) \rightarrow \forall y\, \varphi(y). \qquad (\mathsf{Ind}\text{-}\Delta_0)$$

Partial truth-definitions are available in $I\Delta_0$: they can be used to show that EA is finitely axiomatizable (cf. Hájek & Pudlák (1998, chap. V)). In working with EA, I will often employ functional expressions for elementary functions,[15] such as $\mathsf{S}(x)$, as unofficial counterparts of the relations just introduced. In practice I work in a definitional extension of EA, assuming in the background the possibility of translating back the unofficial abbreviations into the official signatures by eliminating terms via (suitably bounded) existential quantifiers.

The formalization of the syntax of first-order theories in EA is carried out without difficulties: the details of one such coding can be found, for instance, in Schwichtenberg and Wainer (2011). Par abus de language, we denote with $\ulcorner e \urcorner$ the formal code of e. We also avail ourselves with symbols for elementary functions such that EA 'proves' the following equations, with φ, ψ $\mathcal{L}_0$-formulas and x_i a (meta)variable of $\mathcal{L}_0$—standing for v_{i_1}:

$$\dot{R}(\ulcorner x_1 \urcorner, \ldots, \ulcorner x_n \urcorner) = \ulcorner R(x_1, \ldots, x_n) \urcorner \quad \mathsf{ng}(\ulcorner \varphi \urcorner) = \ulcorner \neg \varphi \urcorner \quad \mathsf{dn}(\ulcorner \varphi \urcorner) = \ulcorner \neg\neg \varphi \urcorner$$
$$\mathsf{and}(\ulcorner \varphi \urcorner, \ulcorner \psi \urcorner) = \ulcorner \varphi \wedge \psi \urcorner \quad \mathsf{all}(\ulcorner v \urcorner, \ulcorner \varphi \urcorner) = \ulcorner \forall v \varphi \urcorner.$$

The operation $\varphi(v), t \mapsto \varphi(t)$ of substituting a 'term' for a free variable in a formula is naturally represented in EA by an elementary formula $\mathsf{sub}(x, y, u, v)$ such that

$$\mathsf{EA} \vdash \forall x, y, u\, \exists! v\, \mathsf{sub}(x, y, u, v)$$
$$\mathsf{EA} \vdash \mathsf{sub}(\ulcorner \varphi(x_i) \urcorner, \ulcorner x_i \urcorner, \ulcorner t \urcorner, y) \rightarrow y = \ulcorner \varphi(t) \urcorner.$$

For notational convenience, I will write sub as if it were a function. We also write $\langle x_1, \ldots, x_n \rangle$ to designate in EA a finite sequence, $(x)_i$ to indicate the i^{th} element of the sequence x, $\mathsf{lh}(x)$ for its length. All these operations correspond to elementary functions and are provably total in EA. I will also extensively employ the following elementary syntactic notions relative to some elementary presented theory T in $\mathcal{L}_0$:

$\mathsf{Fml}_{\mathcal{L}_0}(x)$	'x is is the code of an $\mathcal{L}_0$-formula'
$\mathsf{Fml}^i_{\mathcal{L}_0}(x)$	'x is is the code of an $\mathcal{L}_0$-formula with i variables free'
$\mathsf{Sent}_{\mathcal{L}_0}(x)$	'x is is the code of an $\mathcal{L}_0$-sentence'
$\mathsf{Ax}_T(x)$	'x is a logical or nonlogical axiom of T'
$\mathsf{Prf}_T(x, y)$	'x is a proof of y in T'.

From the canonical proof predicate $\mathsf{Prf}_T(x, y)$ one defines the provability predicate $\mathsf{Pr}_T(x)$ as $\exists y\, \mathsf{Prf}_T(x, y)$, expressing that there is a proof (a sequence of axioms or formulas obtained from the axioms by applications of the rules of inference) of x in T. The

[15] The class of *elementary functions* $\mathcal{E}$ is obtained by closing the initial functions $\mathsf{zero}(\cdot)$, $\mathsf{suc}(\cdot)$, $+$, $\times$, 2^x, $\mathsf{P}^n_i(x_1, \ldots, x_n) = x_i$ with $(1 \leq i \leq n)$, truncated subtraction $x \dot{-} y$ under the operations of composition and bounded minimalization:

$$H(\vec{x}) = F(G_1(\vec{x}), \ldots, G_n(\vec{x})); \qquad (\mu t \leq y)\ P(\vec{x}, t) = \begin{cases} \text{the least } t \leq y \text{ s.t. } P(\vec{x}, t) \\ 0, \quad \text{if there is no such } t \end{cases}$$

where F, G are elementary functions and P an elementary predicate. EA has sufficient resources to naturally introduce new relations corresponding to the elementary functions by proving their defining equations.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Π_1-sentence $\mathsf{Con}(T)$, expressing an intensionally correct consistency statement for T, is then simply $\mathsf{Con}(T) :\leftrightarrow \neg \mathsf{Pr}_T(\bot)$, where $\bot$ codes a falsity in T.

In this paper we will mostly deal with the language $\mathcal{L}$ of EA expanded with one (or more) truth predicates and with extensions T of EA with axioms governing the behaviour of these predicates. In formulating the truth axioms in a relational setting, it is useful to resort to the fact that EA can represent the language '$\mathcal{L}$ plus domain constants' via an injection $x \mapsto c_x$ formally associating each object x in the sense of T with these new constants. We will still call $\mathcal{L}$ this expanded formal language internally represented in $\mathcal{L}$ itself and write $\overline{x}$ to denote these formal objects; the new formal constants will enable us to quantify into Gödel corners in the usual way. We write $\ulcorner \varphi(\dot{x}) \urcorner$ or $\mathsf{sub}(\ulcorner \varphi(v) \urcorner, \overline{x})$ for $\mathsf{sub}(\ulcorner \varphi(v) \urcorner, \ulcorner v \urcorner, \overline{x})$.

A sound translation function $\tau : \mathcal{L}_P (:= \mathcal{L} \cup \{P\}) \to \mathcal{L}_1$, applied to sentences of the form $P\ulcorner Pt \urcorner$ and replacing $P(\cdot)$ with some $\mathcal{L}_1$-formula $\xi(\cdot)$, should of course yield $\xi(\dot{\tau}\ulcorner Pt \urcorner)$ and not $\xi(\ulcorner Pt \urcorner)$, where the notation $\dot{\tau}(\cdot)$ refers to a functional expression in $\mathcal{L}$ representing τ in EA. To achieve the required translation, one may resort to the recursion theorem (see §11.2 of Rogers (1987)) that applies equally well to elementary functions; it yields for any recursive $f(x, y)$ an index e such that $f(e, y) \cong \phi_e(y)$, where $\phi_{(\cdot)}(\cdot)$ is the universal program. We can then define an elementary translation function τ_0 such that, in the relevant case, $\tau_0(x, P\ulcorner Pt \urcorner) = \xi(\dot{\phi}_x(\ulcorner Pt \urcorner))$ and apply the recursion theorem to find an index e for τ_0 such that $\phi_e(P\ulcorner Pt \urcorner) = \dot{\xi}(\dot{\phi}_e(Pt))$. We can then let $\tau(x) \cong \phi_e(x)$.

When not otherwise specified, EA will be the base theory for our theories of truth. I do not attempt at giving sufficient conditions for what counts for a theory of truth, although I will often employ this expression. I hope that my choices will be self-explanatory and at any rate they are based on widely accepted choices.

I will employ different measures of the complexity of formulas: the *positive complexity* $|\varphi|^+$ of a formula φ is 0 for atomic and negated atomic formulas; it is $|\psi|^+ + 1$ if φ is $\neg\neg\psi$ or $\forall x \psi$, it is $\mathsf{max}(|\chi|^+, |\psi|^+) + 1$ if φ is $\chi \wedge \psi$. The *logical complexity* of a formula φ—formalized as $\mathsf{lc}(\ulcorner \varphi \urcorner)$—is simply the number of its logical symbols. These measures of complexity correspond to elementary functions and can be naturally represented in EA.

2.2. Interpretations and isomorphisms. The notion of relative interpretation and its variants will be central in this work, so they deserve careful introduction. We will consider only relational languages with finite signatures.

A *relative translation* of $\mathcal{L}_T$ into $\mathcal{L}_W$ can be described as a pair (δ, F) where δ is a $\mathcal{L}_W$-formula with one free variable—the domain of the translation—and F is a (finite) mapping that takes n-ary relation symbols of $\mathcal{L}_T$ and gives back formulas of $\mathcal{L}_W$ with n free variables. The translation extends, modulo suitable renaming of bound variables,[16] to the mapping τ:

- $(R(x_1, \ldots, x_n))^\tau :\leftrightarrow F(R)(x_1, \ldots, x_n)$;
- τ commutes with propositional connectives;
- $(\forall x\ \varphi(x))^\tau \ :\leftrightarrow\ \forall x\ (\delta(x) \to \varphi^\tau)$.

DEFINITION 2.1. *An (one-dimensional) interpretation K is specified by a triple (T, τ, W), where τ is a translation of $\mathcal{L}_T$ in $\mathcal{L}_W$, such that for all formulas $\varphi(x_1, \ldots, x_n)$ of $\mathcal{L}_T$ with the free variables displayed, we have*:

[16] For more details on how clashes are avoided, I refer to (Visser, 1997).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

$$\text{if } T \vdash \varphi(x_1, \ldots, x_n), \text{ then } W \vdash \bigwedge_{i=1}^{n} \delta_K(x_i) \to \varphi^\tau.$$

I write $K : T \to W$ for 'K is an interpretation of T in W'. An interpretation is *direct* if it maps identity to identity and it does not relativize quantifiers. I will often not distinguish between an interpretation and the relative translation that supports it. T and W are said to be *mutually interpretable* if there are interpretations $K : T \to W$ and $L : W \to T$. Since interpretability is a partial preorder on theories, mutual interpretability is an equivalence relation (its equivalence classes are known as degrees of interpretability).

I will often refer to a useful, model-theoretic characterization of interpretability. If $K : T \to W$ and $\mathcal{M}$ is any model of W, then K can be seen as a method for defining a model $\mathcal{M}^K$ of T inside $\mathcal{M}$.

LEMMA 2.2. *If $K : T \to W$ is identity preserving and W has full induction, then for any $\mathcal{M} \models W$ we find a (uniformly) $\mathcal{M}$-definable embedding of $\mathcal{M}$ into an initial segment of $\mathcal{M}^K$.*[17]

Lemma 2.2 is obtained by noticing that in $\mathcal{M}$ we can define an injection $f : \mathcal{M} \to \mathcal{M}^K$ such that, again by employing a convenient functional notation,

$$f(0^{\mathcal{M}}) \mapsto 0^{\mathcal{M}^K}; \qquad f(x +^{\mathcal{M}} 1^{\mathcal{M}}) \mapsto f(x) +^{\mathcal{M}^K} 1^{\mathcal{M}^K}$$
$$f(x +^{\mathcal{M}} y) \mapsto f(x) +^{\mathcal{M}^K} f(y) \qquad f(x \times^{\mathcal{M}} y) \mapsto f(x) \times^{\mathcal{M}^K} f(y).$$

Full induction in $\mathcal{M}$ is needed to prove the totality of f. f is clearly an isomorphism of $\mathcal{M}$ in $\mathcal{M}^K$.

Given an intensionally correct consistency statement $\mathsf{Con}(T)$, for T elementary presented,[18] one can construct a formalized Henkin model of T in $\mathsf{EA}{+}\mathsf{Con}(T)$: the model is constructed in the standard way by means of Henkin axioms and it is equipped with a truth predicate—though not a truth predicate working for *all sentences* in the sense of T:

LEMMA 2.3. *Let the axiom set of T be captured by an elementary predicate. Then, $\mathsf{EA}{+}\mathsf{Con}(T)$ interprets T.*

To introduce stronger notions of interpretations, I introduce compositions of interpretations. Given $\tau_0 : \mathcal{L}_T \to \mathcal{L}_W$ and $\tau_1 : \mathcal{L}_W \to \mathcal{L}_V$, the composite of $K = (T, \tau_0, W)$ and $L = (W, \tau_1, V)$ is the interpretation $L \circ K = (T, \tau_1 \circ \tau_0, V)$, where $\delta_{L\circ K}(x) :\leftrightarrow \delta_K^L(x) \wedge \delta_L(x)$. Two interpretations $K_0, K_1 : T \to W$ are *equal* if W, the target theory, proves this. In particular, one requires,

$$W \vdash \forall x\, (\delta_{K_0}(x) \leftrightarrow \delta_{K_1}(x))$$
$$W \vdash \forall \vec{x}\, (R_{K_0}(\vec{x}) \leftrightarrow R_{K_1}(\vec{x})) \qquad \text{for any relation symbol } R \text{ of } \mathcal{L}_T$$

A W- definable *morphism* between (again, one-dimensional) interpretations $K_0, K_1 : T \to W$ is a triple (K_0, ϕ, K_1), with ϕ a $\mathcal{L}_W$-formula with two free variables, satisfying:

$$W \vdash \forall x, y\, \big(\phi(x, y) \to (\delta_{K_0}(x) \wedge \delta_{K_1}(y))\big) \tag{2}$$
$$W \vdash \forall x, y, u, v \big(x =_{K_0} y \wedge u =_{K_1} v \wedge \phi(y, u) \to \phi(x, v)\big) \tag{3}$$
$$W \vdash \forall x\, \big(\delta_{K_0}(x) \to \exists y\, (\delta_{K_1}(y) \wedge \phi(x, y))\big) \tag{4}$$

[17] The condition on identity is redundant although not obviously so.

[18] The same holds even if we allow T to be presented by a p-time set of axioms. A proof by Albert Visser of this stronger claim can be found in Visser (1991). See also Nicolai (2016a).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

$$W \vdash \forall x, y, z\big(\phi(x, y) \wedge \phi(x, z) \rightarrow y =_{K_1} z\big) \tag{5}$$

$$W \vdash \forall x_1, \ldots, x_n \forall y_1, \ldots, y_n \Big(\bigwedge_{i=1}^{n} \phi(x_i, y_i) \wedge R_{K_0}(x_1, \ldots, x_n) \rightarrow R_{K_1}(y_1, \ldots, y_n)\Big)\Big) \tag{6}$$

for any relation $R \in \mathcal{L}_T$.

To obtain an *isomorphism* from K_0 to K_1 one needs to add the following, extra conditions:

$$W \vdash \forall y \left(\delta_{K_1}(y) \rightarrow \exists x \left(\delta_{K_0}(x) \wedge \phi(x, y)\right)\right) \tag{7}$$

$$W \vdash \forall x, y, z \left(\phi(x, y) \wedge \phi(z, y) \rightarrow x =_{K_0} z\right) \tag{8}$$

$$W \vdash \forall x_1, \ldots, x_n \forall y_1, \ldots, y_n \Big(\bigwedge_{i=1}^{n} \phi(x_i, y_i) \wedge R_{K_1}(y_1, \ldots, y_n) \rightarrow R_{K_0}(x_1, \ldots, x_n)\Big)\Big) \tag{9}$$

for any relation $R \in \mathcal{L}_T$.

We write $F\colon K_0 \cong K_1$ for 'F is an isomorphism from the interpretation K_0 to K_1'.

2.3. Sameness of theories. In this subsection I introduce two noticeable notions of sameness of theories extending extensional identity: bi-interpretability and synonymy (cf. Visser (2006), Friedman & Visser (2014)). They will play an important role in what follows.

DEFINITION 2.4 (Synonymy). *U and V are synonymous if and only if there are interpretations $K\colon U \rightarrow V$ and $L\colon V \rightarrow U$ such that V proves that $K \circ L$ and id_V are equal and, symmetrically, U proves that $L \circ K$ is equal to id_U.*

One can check that for U and V formulated in disjoint signatures, U and V are synonymous if and only if they have a common definitional extension. As shown in Kaye & Wong (2007), for instance, ZF minus the axiom of infinity plus its negation and 'every set has a transitive closure' is synonymous with PA.[19]

To introduce a slightly less strict notion of equivalence, *bi-interpretability*, we first consider the notion of a retract, that in turn relies on the notion of isomorphism of interpretations introduced in the previous section.

DEFINITION 2.5 (Retract). *U is a retract of V if and only if there are $K\colon U \rightarrow V$ and $L\colon V \rightarrow U$ and a U-definable isomorphism between $L \circ K$ and id_U.*

The category-theoretic lexicon is due to Visser's systematization of various notions of equivalence of theories in terms of appropriate categories of theories and interpretations (Visser, 2006). We refer to Visser's outstanding paper for further insights on the category-theoretic presentation. The notion of retract can also be visualized in model-theoretic terms: as it is illustrated in Figure 1, when $K\colon U \rightarrow V$ and $L\colon V \rightarrow U$, U is a retract of V if in any model $\mathcal{M} \models U$ we can construct an internal $\mathcal{M}^L \models V$ which in turn defines a model $(\mathcal{M}^L)^K \models U$ and there is an $\mathcal{M}$-definable isomorphism $F\colon (\mathcal{M}^L)^K \cong \mathcal{M}$.

Essentially, U and V are bi-interpretable if and only if U is a retract of V and V is a retract of U using the same pair of interpretations.

[19] ZF minus infinity plus its negation, rather surprisingly, is neither synonymous nor bi-interpretable with PA, (cf. Enayat, Schmerl, & Visser (2010)).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

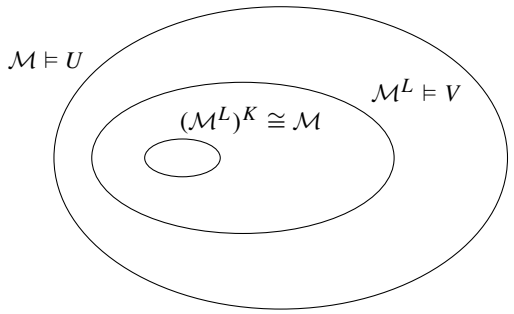


Fig. 1. U is a retract of V.

DEFINITION 2.6 (Bi-interpretability). *Given a pair of interpretations $K\colon U \to V$ and $L\colon V \to U$, U and V are bi-interpretable if and only if (i) there is a $\mathcal{L}_V$-formula F_0 such that V proves F_0 to be an isomorphism between $K \circ L$ and id_V and (ii) there is an $\mathcal{L}_U$-formula F_1 such that U proves F_1 to be an isomorphism between $L \circ K$ and id_U.*

Clearly, if U and V are synonymous, they are also bi-interpretable. The converse is, provably, not true. In Friedman & Visser (2014) it is shown, for instance, that Adjunctive Set theory and its two-sorted, flattened version with a Frege function are bi-interpretable but not synonymous. Bi-interpretability preserves many mathematical properties of the theories: decidability, finite axiomatizability (cf. Lemma 2.8), κ-categoricity. That synonymy is a stricter notion than bi-interpretability follows from the fundamental insight that the latter does preserve the automorphism group of models modulo isomorphism but not the action of the automorphism group on the domain of the model. Synonymy preserves both.

An important fact linking synonymy and bi-interpretability is the following, involving the so-called *sequential* theories.[20]

LEMMA 2.7 (Friedman & Visser, 2014). *Let U, V be sequential. If $K\colon U \to V$ and $L\colon V \to U$ witness a bi-interpretation of U and V and L is unrelativized and identity preserving, then U and V are synonymous.*

As hinted above, synonymy and interpretability can be seen, respectively, as equality and isomorphism in appropriate categories INT_0 and INT_1 of theories and interpretations (cf. Visser (2006)). A category-theoretic framework is arguably the best way to formulate a general theory of inter-theoretic reductions: in this paper, however, we will only employ the category-theoretic jargon as convenient notation. Since it will be relevant in what follows, we conclude the subsection with a sketch of the proof that bi-interpretability (and thus, synonymy), preserves finite axiomatizability.

LEMMA 2.8 (Visser, 2006). *Let U, V be theories in finite signatures. Assume that $K\colon U \to V$ and $L\colon V \to U$ are interpretations and that U defines an isomorphism F from $L \circ K$ to id_U. Assume further that V is finitely axiomatizable. Then U is finitely axiomatizable.*

Proof sketch. Let V_0 be the conjunction of a finite axiomatization of V. A finite $U_0 \subseteq U$ is specified by the single sentences: (i) F is an isomorphism between $L \circ K$ and id_U;

[20] A theory is sequential if it directly interprets Adjunctive set theory. More informally, a theory is sequential is it has a good coding of sequences. See Friedman & Visser (2014) for details.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

(ii) V_0^L. The theory U_0 is, by definition of retract, a subtheory of U. For the converse direction, one verifies that if U proves the sentence φ, then $U_0 \vdash \varphi^{KL}$ by (ii) and the definition of retract. Thus $U_0 \vdash \varphi$ by (i). □

§3. The usual suspects. One option to formally capture the notion of conceptual equivalence of truth predicates may be to consider the *truth-free consequences* of the theories that characterize them. This is notoriously a bad idea. For my purposes here, it is useful to support this claim by introducing two theories that will be employed also in later sections.

DEFINITION 3.1. *The theory* EA^{T} *in* $\mathcal{L}_{\mathsf{T}} := \mathcal{L} \cup \{\mathsf{T}\}$ *is obtained by formulating* EA *in the new language and allowing the truth predicate into instances of* Δ_0-Ind. EA^{t}*, in turn, only features arithmetical* Δ_0*-induction.*

DEFINITION 3.2 (TB). *The theory* TB *is obtained by extending* EA^{T} *with all instances of the schema*

$$\mathsf{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi \tag{tb}$$

for φ *a sentence of* $\mathcal{L}$. $\mathsf{TB}\restriction$ *is obtained by adding* (tb) *to* EA^{t}.

The other theory is essentially the axiom set introduced by Feferman and inspired to the Strong Kleene version of Kripke's theory of truth (Cantini, 1989; Feferman, 1991).

DEFINITION 3.3 (KF$\restriction$). $\mathsf{KF}\restriction$ *extends* EA^{t} *with the universal closures of the following*:

$$(\mathsf{T}\ulcorner R(\dot{x}_1, \ldots, \dot{x}_n)\urcorner \leftrightarrow R(x_1, \ldots, x_n)) \wedge (\mathsf{Tng}(\ulcorner R(\dot{x}_1, \ldots, \dot{x}_n)\urcorner) \leftrightarrow \neg R(x_1, \ldots, x_n)) \quad \mathsf{Tat}$$

$$\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(x) \rightarrow \big(\mathsf{Tdn}(x) \leftrightarrow \mathsf{T}x\big) \quad \mathsf{T1}$$

$$\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(\mathsf{and}(x, y)) \rightarrow \big(\mathsf{Tand}(x, y) \leftrightarrow \mathsf{T}x \wedge \mathsf{T}y\big) \quad \mathsf{T2}$$

$$\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(\mathsf{and}(x, y)) \rightarrow \big(\mathsf{Tng}(\mathsf{and}(x, y)) \leftrightarrow \mathsf{Tng}(x) \vee \mathsf{Tng}(y)\big) \quad \mathsf{T3}$$

$$\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(\mathsf{all}(v, x)) \rightarrow \big(\mathsf{Tall}(v, x) \leftrightarrow \forall y\, \mathsf{Tsub}(x, \overline{y})\big) \quad \mathsf{T4}$$

$$\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(\mathsf{all}(v, x)) \rightarrow \big(\mathsf{Tng}(\mathsf{all}(v, x)) \leftrightarrow \exists y\, \mathsf{Tsub}(\mathsf{ng}(x), \overline{y})\big) \quad \mathsf{T5}$$

$$\mathsf{T}\ulcorner\mathsf{T}\dot{x}\urcorner \leftrightarrow \mathsf{T}x \quad \mathsf{T6}$$

$$\mathsf{Tng}(\ulcorner\mathsf{T}\dot{x}\urcorner) \leftrightarrow \mathsf{Tng}(x) \quad \mathsf{T7}$$

$$\mathsf{T}x \rightarrow \mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(x) \quad \mathsf{T8}$$

The arguments that witness the conservativeness of $\mathsf{TB}\restriction$ and $\mathsf{KF}\restriction$ over EA are well-known. For $\mathsf{TB}\restriction$ one replaces in every derivation in $\mathsf{TB}\restriction$ of a truth-free formula the truth predicate with an EA-definable, partial truth predicate. In the second case, starting with any model of EA (in the language with domain constants), we expand it via a positive inductive definition of a Kripke truth set to a model of $\mathsf{KF}\restriction$ (see §4 of Cantini (1989)).

LEMMA 3.4. $\mathsf{TB}\restriction$, EA, *and* $\mathsf{KF}\restriction$ *prove the same* $\mathcal{L}$*-theorems.*

By contrast, it is not hard to see that $\mathsf{KF}\restriction$ and$\mathsf{TB}\restriction$ fare much differently with respect to the conditions in §2. $\mathsf{TB}\restriction$ is typed and $\mathsf{KF}\restriction$ is type-free; $\mathsf{KF}\restriction$ is compositional to a large extent—only negation, for obvious reasons, does not fully commute with T—whereas $\mathsf{TB}\restriction$ can hardly be thought as compositional: for instance, $\mathsf{TB}\restriction$ cannot prove the sentence

$$\forall x, y\, (\mathsf{Sent}_{\mathcal{L}}\big(\mathsf{and}(x, y)\big) \rightarrow (\mathsf{T}(\mathsf{and}(x, y)) \leftrightarrow (\mathsf{T}x \wedge \mathsf{T}y))\big). \tag{10}$$

The unprovability of (10) also suggests that TB falls short of many generalizations provable in $\mathsf{KF}\restriction$, suggesting that also this desideratum separates the two theories. Finally, TB and

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

$\mathsf{KF}\restriction$ are theories whose truth predicates capture, in the sense of Fischer et al. (2015) instantiated by (1), different semantic constructions: the Tarskian truth set and the fixed points of Kripke's theory of truth respectively.

The case just considered is not isolated: It is not hard to find many other counterexamples to the claim that the provability of the same truth-free consequences amounts to an adequate formal rendering of the conceptual equivalence of two theories' truth predicates.

An alternative may be represented by *mutual interpretability*. A little reflection shows, however, that mutual interpretability suffers from a problem that is in a sense opposite to the one suffered by provability of the same base-theoretic consequences: it mixes up syntactic and semantic aspects of our theories.

LEMMA 3.5 (Feferman). *Any* $T \supseteq \mathsf{EA}$ *interprets* $T + \neg\mathsf{Con}(T)$.

Proof sketch. This proof is due to Visser and, independently, to Lindström; it does not rely on the reflexivity of the theory involved and therefore suits our purposes. By Gödel's second incompleteness theorem, $T \vdash \mathsf{Con}(T) \rightarrow \mathsf{Con}(T + \neg\mathsf{Con}(T))$, therefore $T + \mathsf{Con}(T)$ interprets $T + \neg\mathsf{Con}(T)$ by Lemma 2.3, let's say with an interpretation K. Finally we combine the identity interpretation on $T + \neg\mathsf{Con}(T)$ and K: if $\mathsf{Con}(T)$, pick K, if $\neg\mathsf{Con}(T)$, $\mathsf{id}_{T+\neg\mathsf{Con}(T)}$ suffices. □

By Lemma 3.5, any theory of truth over EA will mutually interpret its inconsistency. Since we usually require syntactic notions—including Con(T)—to be canonically constructed and T to be ω-consistent, any theory of truth will be 'equivalent' with a theory that compromises our basic assumptions on the structure of syntactic universe.[21] Mutual interpretability does not guarantee that the structure of the so-called bearers of truth remains fixed across theories. For the reasons explained in §2, this seems to undermine the very possibility of comparing truth predicates.[22]

To overcome these and further difficulties with assuming provability of the same base-theoretic consequences or mutual interpretability as formal counterparts of conceptual equivalence of two theories of truth, we may resort to the notion of *relative truth-definability* due to Fujimoto (2010) and Halbach (2014). In a nutshell, T is truth-definable in W if there is a direct, relative interpretation of T in W that preserves the vocabulary of the syntactic base theory B. I give a more pedantic definition to fix some notation, starting with an arbitrary base theory B containing EA. It is convenient here to consider the language of our theories of truth as extending the base language $\mathcal{L}_B$ with an indexed set of semantic predicates $\{S_i\}_{i\in I}$: we want to allow in fact for the possibility that truth is expressed via, for instance, a binary satisfaction predicate.

---

[21] An advocate of mutual interpretability, at this point, may highlight the artificiality of the theories like $\mathsf{T} + \neg\,\mathsf{Con}(T)$. After all, they are artifacts just right to reach Σ_1-unsoundness, therefore ω-inconsistency. However, more natural examples are not difficult to construct. To mention a familiar example, let us consider the theory $\mathsf{TB}[\mathsf{PA}]$ just introduced and the theory resulting from extending PA with restricted induction with Tat, the typed versions of $\mathsf{T2}$, $\mathsf{T4}$ and

$$\mathsf{Sent}_{\mathcal{L}}(x) \rightarrow \big(\mathsf{Tng}(x) \leftrightarrow \neg\,\mathsf{T}x\big), \tag{11}$$

known in the literature as $\mathsf{CT}\restriction[\mathsf{PA}]$ or $\mathsf{PA}^{\mathsf{FT}}$. $\mathsf{TB}[\mathsf{PA}]$ and $\mathsf{CT}\restriction[\mathsf{PA}]$ are mutually interpretable, as $\mathsf{CT}\restriction[\mathsf{PA}]$ is interpretable in PA (cf. Enayat & Visser (2015)), but the theories fare rather differently in terms of the conditions in §2 and do not share the same conception of truth in the sense defined above.

[22] Fujimoto (2010) already employed Feferman's theorem on the interpretability of inconsistency—applied to TB—to argue against mutual interpretability as a notion of theoretical equivalence for theories of truth. For a recent, thorough study of Feferman's theorem, see Visser (2015).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

DEFINITION 3.6 (Truth translation). *A* truth-translation $\tau\colon \mathcal{L}_T \to \mathcal{L}_W$, *where* $\mathcal{L}_T = \mathcal{L}_B \cup \{S_i\}_{i\in I}$ *and* $\mathcal{L} \in \mathcal{L}_T \cap \mathcal{L}_W$, *is a pair* (d, F), *where* $\mathsf{d}(x)$ *is* $x = x$ *and* F *a mapping of* n*-ary predicates of* $\mathcal{L}_T$ *into* $\mathcal{L}_W$ *formulas with* n *free variables satisfying*:

$$\begin{aligned}
\tau(R)(\vec{x}) &:\leftrightarrow R(\vec{x}) && \text{for } n\text{-ary } R \in \mathcal{L}_B;\\
\tau(S_i)(\vec{x}) &:\leftrightarrow F(S_i)(\vec{x}) && \text{for } i \in I\\
\tau(\neg\varphi) &:\leftrightarrow \neg\tau(\varphi)\\
\tau(\varphi \wedge \psi) &:\leftrightarrow \tau(\varphi) \wedge \tau(\psi)\\
\tau(\forall x\varphi) &:\leftrightarrow \forall x(\mathsf{d}(x) \to \tau(\varphi)).
\end{aligned}$$

The mapping F behaves like the identity function when applied to base-theoretic, nonlogical symbols. Moreover, the translation is clearly unrelativized.

DEFINITION 3.7 (Relative truth-definition). *A* truth-definition *of a theory* T *in* $\mathcal{L}_T = \mathcal{L}_B \cup \{S_i\}_{i\in I}$*—with* I *an index set—in a theory* W *is a triple* (T, τ, W), *where* τ *is a truth-translation of* $\mathcal{L}_T$ *into* $\mathcal{L}_W$ *and for all* $\mathcal{L}_T$*-sentences* φ, *if* $T \vdash \varphi$, *then* $W \vdash \tau(\varphi)$.

Like relative interpretability, relative truth-definability is a partial preorder. We call two theories of truth T and W *mutually truth-definable* if there are truth-definitions $K\colon T \to W$ and $L : W \to T$. Mutual truth-definability defines indeed an equivalence relation on theories. As Fujimoto (2010) points out, it is immediate from the definition that if W truth-defines T and any $\mathcal{M} \vDash B$ admits an expansion to a model of W, then $\mathcal{M}$ admits an expansion to a model of T.[23] Also, due to the non-relativization of quantifiers, if T is Σ_1-unsound or ω-inconsistent and T is truth-definable in W, W will be Σ_1-unsound or ω-inconsistent. Therefore, mutual truth-definability prevents the possibility of equating theories with non-isomorphic ontological commitments. Also, it takes care of what the theory proves true as if T truth defines W, there should be a formula of the language of T defining the extension of the truth predicate of W—in the case of a single truth predicate in W—simulating the behaviour of this truth predicate. But is this enough to conclude that the truth predicates of mutually truth-definable theories embody the same conception of truth? There are well-known examples, it seems, that suggest a negative answer to this question. We consider some of them.

In the following a formula of $\mathcal{L}_{\mathsf{T}}$ will be said to be T-*positive* if and only if T does not occur, in φ, in the scope of an odd number of negation signs. Halbach (2009) introduced and motivated a collection of type-free, uniform T-sentences that escape the Liar paradox by controlling the behaviour of negation.

DEFINITION 3.8. PUTB *is the theory obtained by extending* EA^{T} *with the schema*

$$\forall x\left(\mathsf{T}(\ulcorner\varphi(\dot{x})\urcorner) \leftrightarrow \varphi(x)\right) \tag{putb}$$

for all T*-positive formulas* φ *of* $\mathcal{L}_T$.

DEFINITION 3.9. KF *is the theory obtained by extending* EA^{T} *with* Tat-T7.

LEMMA 3.10. PUTB *is a subtheory of* KF.

[23] This is simply because the numbers of the two models are the same, and the truth-definition ensures that the right extension of the truth predicates of T can be always extracted via the extension of the truth predicate of W.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Lemma 3.10 is obtained by showing, via a meta-induction on the positive complexity of $\mathcal{L}_{\mathsf{T}}$-formulas, that (putb) is provable in KF. Trivially, therefore, KF defines the truth predicate of PUTB. More interestingly, Halbach (2009) also shows that

LEMMA 3.11. PUTB *defines the truth predicate of* KF.

Proof sketch. EA^{T} proves the diagonal lemma in parametrized form, i.e.,: for any $\mathcal{L}_{\mathsf{T}}$-formula $\varphi(x, \vec{y})$, we find a unary formula $\psi(\vec{y})$ such that

$$\mathsf{EA}^{\mathsf{T}} \vdash \varphi(\vec{y}) \leftrightarrow \psi(\ulcorner \varphi(\vec{y}) \urcorner, \vec{y}). \tag{12}$$

Now if we let $\mathcal{I}(x, \mathsf{T}y)$ be a $\mathcal{L}_{\mathsf{T}}$-formula that mimics the positive inductive definition of a Kripke truth-set under the scope of T (see p. 190 of Halbach (2014)), (12) will yield a formula $\theta(x)$ such that

$$\mathsf{EA}^{\mathsf{T}} \vdash \theta(x) \leftrightarrow \mathcal{I}(x, \mathsf{T}\ulcorner \theta(x) \urcorner). \tag{13}$$

Since in $\theta(x)$ there are only positive occurrences of T, they can be removed so that one can verify that $\theta(x)$ satisfies all axioms of KF. □

If mutual truth-definability were a sufficient condition to establish the conceptual equivalence of the theories' truth predicates, the concepts of truth associated with PUTB and KF would be equivalent. This is the position taken, for instance, by Fujimoto (2010).[24] Despite their mutual truth-definability, PUTB and KF fare much differently with respect to the criteria highlighted in §2. Firstly, PUTB is can hardly be considered to be *compositional* and it cannot prove as many general claims as KF by using its own notion of truth. As shown in Halbach (2009, Lemma 6.1), in fact, any proof in PUTB only employs positive disquotational sentences applied to $\mathcal{L}_{\mathsf{T}}$-sentences φ with limited, standard complexity. But there is more: PUTB can be consistently made symmetric, that is the result of closing PUTB under the rules

$$\text{(T-in)} \quad \frac{\varphi(x)}{\mathsf{T}\ulcorner \varphi(\dot{x}) \urcorner} \qquad\qquad \text{(T-out)} \quad \frac{\mathsf{T}\ulcorner \varphi(\dot{x}) \urcorner}{\varphi(x)}$$

for $\varphi(v) \in \mathcal{L}_{\mathsf{T}}$, yields a consistent theory.[25] It is in fact possible to rule out putative proofs of inconsistencies in PUTB plus (T-in) and (T-out) in the following way: if (T-out) and

[24] He writes:

> ...[the mutual truth-definability of KF and PUTB] may be still interpreted to indicate that KF and PUTB are indeed 'conceptually equivalent' despite superficial differences. (Fujimoto, 2010, p. 342)

To be fair with Fujimoto's position, he also claims:

> ...we propose [relative truth-definability, A/N] 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). Fujimoto (2010)

The issue thus becomes dangerously verbal. At any rate, even if we take mutual truth-definability to witness a weak form of 'conceptual equivalence', what I will say in what follows may be seen as an invitation to investigate and an attempt to regiment stronger notions of conceptual equivalence.

[25] Of course, to avoid the Liar paradox, (T-in) and (T-out) need to be understood as rules applying only to *theorems* and not to assumptions.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

(T-in) are applied to T-positive sentences they can be dispensed with via (putb). Moreover, in any proof in PUTB plus (T-in) and (T-out) — where the two rules are applied to sentences that are not T-positive — (i) applications of (T-out) may be eliminated; (ii) the semantics of PUTB can be adapted to accommodate possible applications of (T-in) (see Halbach (2014), §19.5, for proofs of these facts). By contrast, KF is essentially asymmetric, as it cannot be consistently closed under the two rules. Finally, if KF is sound with respect to fixed points of the four-valued version of Kripke's semantic construction, PUTB is not as it has models of the form $(\mathbb{N}, S)$ where S is not such a fixed point.

Mutual truth-definability, therefore, does not preserve compositionality (in the sense specified above), provable generalizations, symmetry, at least if one takes at face value *the* truth predicate(s) of the theories. Moreover, mutually truth-definable theories may not be equally close to a semantic construction. In the next section we will see that not even finite axiomatizability is preserved. Back to our example, there is no doubt that PUTB has the resources to replicate the behaviour of the truth predicate of KF via a non-primitive notion, but this would hardly count as evidence for the claim that PUTB and KF are conceptually equivalent. For this matter, only the primitive notion of truth of the theories should matter.

§4. Notions of equivalence of truth predicates. In this section I consider the varieties of equivalences of theories introduced in §2 and investigate whether they may help in overcoming the problems encountered above. My strategy will be suitably adapting those definitions to obtain notions that

(I) contain mutual truth-definability—that is, the latter notion will play the role of necessary condition for the new notions;

(II) properly extend mutual truth-definability so to capture the differences, for instance, between the truth predicates of PUTB and KF;

(III) are non-empty, in the positive sense that there are natural theories of truth that fall into the relations.

Since this is in a sense unexplored territory, it is not my main focus to find notions of inter-theoretic reduction that are *just right* to capture conceptual equivalence of truth predicates: by contrast, I look for strict notions that may eventually be relaxed and calibrated to accommodate our philosophical purposes. In practice, I will proceed as follows: the notions of retract, bi-interpretability and synonymy from §2.2 are, by definition, not in continuity with mutual truth-definability. This is simply because they are not fine tuned for theories extending a common syntax theory with truth axioms: that is, they do not require the interpretations witnessing them to be unrelativized and behaving like the identity on the syntactic vocabulary. There may be no such a thing as syntactic vocabulary in them. Therefore, I modify the definitions of retract and bi-interpretation accordingly by requiring the composed interpretations to be truth-definitions. I will also not forget about the notion of bi-interpretability simpliciter and consider examples of theories of truth that are bi-interpretable but not mutually truth-definable.

DEFINITION 4.1 (t-retract). *Let T, W be theories of truth based on a syntactic base theory B. T is a t-retract of W if there are truth-definitions $K : T \to W$ and $L : W \to T$ and a T-definable isomorphism between $L \circ K$ and id_T.*

At the intuitive level, if T is a *t-retract* of W, then W functions as a faithful mirror for T: the image of T reflected via its truth-definition in W is a faithful one modulo isomorphism. There is no guarantee, however, that also the converse is true: it may well be the case

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

that, when W looks for its image reflected in T via its truth-definition in W, the result is a distorted image. The notion of *t-equivalence*, by contrast, requires both theories to be 'faithful mirrors' focusing on the same pairs of truth-definitions:

DEFINITION 4.2 (t-equivalence). *Let T, W be theories of truth based on B. T and W are truth equivalent iff there are truth-definitions $K\colon T \to W$ and $L\colon W \to T$ such that (i) T proves that there is an isomorphism between $L \circ K$ and id_T; (ii) W proves that there is an isomorphism between $K \circ L$ and id_W.*

It is immediate from the definitions that *t-retractions* and *t-equivalences* simply impose further conditions on mutual truth-definitions. Focusing on them, the fulfillment of desideratum (I) on page 16 is then immediately obtained.

LEMMA 4.3. *If T is a t-retract of W, then T and W are mutually truth-definable.* A fortiori*, if T and W are t-equivalent, then they are mutually truth-definable.*

Lemma 2.7 tells us that, since we are dealing with direct interpretations, t-equivalence is a form of synonymy. I therefore propose the following, quasi-formal thesis:

THESIS 1. *If two theories are t-equivalent, their associated semantic notions are conceptually equivalent.*

Thesis 1 only imposes sufficient conditions on the conceptual equivalence of semantic notions. Obviously, it is not a formally precise claim, although its plausibility seems to be safely grounded in the strictness of the notion of synonymy. Nonetheless, it requires a suitable amount of empirical data to be confirmed or dismissed: in the following sections I start collecting these data by showing that Thesis 1 captures the non-conceptual equivalence of KF and PUTB and that there are simple, reassuring cases of t-equivalence.

4.1. Separating mutual truth-definability, t-retracts, t-equivalence. In this section I consider desideratum (II) on page 16 and show that the notions of t-retract and t-equivalence are properly stricter than mutual truth-definability in a formally precise sense.

PROPOSITION 4.4. PUTB *is not finitely axiomatizable.*

Proof. Let's assume that there is a sentence A such that $\mathsf{A} \dashv\vdash \mathsf{PUTB}$. By the finite axiomatizability of EA, A has the normal form $\mathsf{A}_0 + \mathsf{A}_1$, where A_0 is a finite set of instances of (putb) and A_1 is a finite version of EA^T. The instances of (putb) in A_0 are therefore applied only to sentences φ such that $\mathsf{lc}(\ulcorner\varphi\urcorner) \leq n$ (the number of logical symbols in φ), for a standard n. By adapting the semantics of PUTB given in Halbach (2009) I show that there is a model of A that is not a model of PUTB.

Let $\mathsf{Sent}^n_{\mathcal{L}^+_\mathsf{T}}$ be the set of T-positive sentences of $\mathcal{L}_\mathsf{T}$ with no more than n logical symbols (excluding the identity symbol), for a suitable $n \in \omega$ and let

$$\mathsf{C} :\leftrightarrow \neg(\underbrace{0 = 0 \wedge \ldots \wedge 0 = 0}_{\wedge \text{ applied } n\text{-times}})$$

Notice that, despite the presence of the negation symbol, C is trivially T-positive. We define the operator $\Phi\colon \mathcal{P}(\omega) \to \mathcal{P}(\omega)$:

$$\Phi(S_1) :=\{\varphi \in \mathsf{Sent}^n_{\mathcal{L}^+_\mathsf{T}} \mid (\mathbb{N}, S_1) \vDash \varphi\} \cup \{m \in S_1 : m \notin \mathsf{Sent}^n_{\mathcal{L}^+_\mathsf{T}}\} \cup \{\mathsf{C}\}.$$

It should be clear that $\Phi(\cdot)$ is monotone. The monotonicity of Φ entails the existence of fixed points of Φ and, in particular, of its minimal fixed point $\mathcal{I}_\Phi$ obtained by closing the empty set under Φ. It should be noticed that $\mathsf{C} \in \mathcal{I}_\Phi$ (and it is in any fixed point of Φ).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Moreover, $(\mathbb{N}, \mathcal{I}_\Phi)$ is a model of A. The axioms of EA clearly hold in $(\mathbb{N}, \mathcal{I}_\Phi)$. For $\varphi \in \mathsf{Sent}^n_{\mathcal{L}^+_\mathsf{T}}$, we have

$$(\mathbb{N}, \mathcal{I}_\Phi) \vDash \mathsf{T}\varphi \Leftrightarrow \varphi \in \mathcal{I}_\Phi \ (= \Phi(\mathcal{I}_\Phi)) \Leftrightarrow (\mathbb{N}, \mathcal{I}_\Phi) \vDash \varphi.$$

To complete the proof of Proposition 4.4: given the assumption of the extensional identity of A and PUTB, $(\mathbb{N}, \mathcal{I}_\Phi)$ would also be a model of PUTB. But then $(\mathbb{N}, \mathcal{I}_\Phi) \vDash \mathsf{T}\ulcorner\mathsf{C}\urcorner$, thus $(\mathbb{N}, \mathcal{I}_\Phi) \vDash \mathsf{C}$. That is, the desired contradiction. □

COROLLARY 4.5. $\mathsf{PUTB}\restriction$ *is not finitely axiomatizable.*[26]

Since, by inspection, the theory that we called KF is finitely axiomatizable (recall that we are constructing KF over EA^T!):

COROLLARY 4.6. *Relative truth-definability does not preserve finite axiomatizability.*

But, by Lemma 2.8, finite axiomatizability is preserved over retractions:

COROLLARY 4.7.

(i) PUTB *is not a retract and therefore not a t-retract of* KF;
(ii) PUTB *and* KF *are not bi-interpretable, therefore also not t-equivalent.*

The argument would still go through if we replace EA^T with EA^t. Therefore, Corollary 4.7 holds if we replace PUTB and KF with $\mathsf{PUTB}\restriction$ and $\mathsf{KF}\restriction$ respectively.

A similar line of reasoning can be applied to other cases, less relevant for our discussion but still worth considering.

DEFINITION 4.8.

(i) $\mathsf{CT}\restriction$ *is the theory in* $\mathcal{L}_\mathsf{T}$ *that extends* EA^t *with* Tat*, the restrictions of* $\mathsf{T2}$*,* $\mathsf{T4}$*,* $\mathsf{T8}$ *to sentences of* $\mathcal{L}$*, and the sentence*

$$\mathsf{Sent}_\mathcal{L}(x) \rightarrow \big(\mathsf{Tng}(x) \leftrightarrow \neg\mathsf{T}x\big) \qquad (\star)$$

(ii) $\mathsf{UTB}\restriction$ *is the theory obtained by extending* EA^t *with the schema*

$$\forall x\, (\mathsf{T}\ulcorner\varphi(\dot{x})\urcorner \leftrightarrow \varphi(x)) \qquad (\mathsf{utb})$$

for all formulas $\varphi(v)$ *of* $\mathcal{L}$ *with one free variable.*

By the finite axiomatizability of EA, it follows that $\mathsf{CT}\restriction$ is finitely axiomatizable. A simplified version of the argument in Proposition 4.4 shows that $\mathsf{UTB}\restriction$ is not finitely axiomatizable.

COROLLARY 4.9. $\mathsf{UTB}\restriction$ *is not a t-retract of* $\mathsf{CT}\restriction$*. Therefore,* $\mathsf{CT}\restriction$ *and* $\mathsf{UTB}\restriction$ *are not t-equivalent.*

---

[26] The method of Proposition 4.4 can be applied more generally to yield the non finite axiomatizability of typed T-sentences over an arbitrary base theory in predicate logic. For instance, axioms of the form

$$\mathsf{T}c_i \leftrightarrow \varphi_i$$

for all $i \in \omega$ are not finitely axiomatizable over predicate logic PRED alone.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Unlike Corollary 4.7, the failure of t-equivalence between $\mathsf{UTB}\restriction$ and $\mathsf{CT}\restriction$ does not amount to a further example separating mutual truth-definability and the stricter notions we are considering. Although $\mathsf{UTB}\restriction$ is trivially truth-definable in $\mathsf{CT}\restriction$, the converse fails.[27]

I conclude this subsection with few remarks on desideratum (II) in the light of the observations above. Corollary 4.7 relies on the finite axiomatizability of KF (and $\mathsf{KF}\restriction$) that, in our setting, is easily reached by the finite axiomatizability of EA. In §2 we included the finite axiomatizability of a theory as a property that should be preserved between theories whose truth predicates are considered to be conceptually equivalent. This is a delicate point and we should pause a little more on it: It would seem in fact that two theories like KF and $\mathsf{KF}[\mathsf{PA}]$—that is $\mathsf{Tat} - \mathsf{T8}$ added to EA^{T} or to its version with full induction—*should* be conceptually equivalent; yet this is ruled out as $\mathsf{KF}[\mathsf{PA}]$ is not finitely axiomatizable. Can this asymmetry threaten the entire project of a formal analysis of conceptual equivalence between truth theories?

This concern can be addressed by elaborating on what has been already anticipated in §2. We are considering comparisons of axiomatic systems, and there is considerable evidence for the claim that theories of truth built on different base theories are in a sense incomparable. What we compare are combinations of truth-theoretic and syntactic principles, where a syntax theory is fixed and 'universal'.[28] If we allowed for a comparison of theories of truth build on non-equiconsistent base theories, for instance, we would violate some of the very criteria on which truth predicates are in general evaluated. $\mathsf{CT}\restriction[\mathsf{PA}]$, for instance, is interpretable in PA; $\mathsf{CT}\restriction$, by contrast, is provably not interpretable in EA. If we let $\mathsf{CT}\restriction[\mathsf{PA}]$ and $\mathsf{CT}\restriction$ share the same conception of truth in the strong sense we are after, we would also be led to consider as conceptually equivalent a truth predicate that can be relatively interpreted in the base theory, and a truth predicate that cannot. Therefore two theories sharing the same conception of truth may be regarded in considerably different ways by, for instance, advocates of the expressive irreducibility of the notion of truth.

After all, it seems, it's a sensible choice to keep the syntax theory fixed. But there seem to be reasons to go even further and motivate the choice of EA as base theory for this particular project. Its finite axiomatizability harmonizes with the choice of our truth axioms in a better way than other choices. By adding a suitable infinite set of truth axioms to EA it may be the case—and it is in the case of TB, UTB and PUTB as defined above—that we obtain a non finitely axiomatizable extension of EA; by adding a finite set of truth axioms to it, like in the case of KF and CT, we obtain a finitely axiomatizable extension of EA. If other choices of the base theories are made, e.g., PA, it is often the case that the size of our set of truth axioms does not matter for the finite axiomatizability of the resulting theory of truth. $\mathsf{TB}[\mathsf{PA}]$, for instance, is not finitely axiomatizable. Despite the finite axiomatizability of $\mathsf{CT}\restriction$, however, $\mathsf{CT}\restriction[\mathsf{PA}]$ is not finitely axiomatizable essentially because of the reflexivity of PA (cf. Fischer (2009), Leigh (2015), Enayat & Visser (2015)).

4.2. t-equivalence, t-retracts, bi-interpretability: some examples. In this section I move to desideratum (III) and introduce theories of primitive truth that are t-retracts of others or that are t-equivalent to others. By definition, t-equivalence satisfies the criteria for Lemma 2.7 and leads straight to synonymy (or definitional equivalence under minimal

[27] Otherwise the theory CT obtained from $\mathsf{CT}\restriction$ by allowing $\mathcal{L}_{\mathsf{T}}$-formulas into the schema of Δ_0-induction will be interpretable in UTB, namely $\mathsf{UTB}\restriction$ with Δ_0-T induction, quod non as CT proves $\mathsf{Con}(\mathsf{EA})$.

[28] By the considerations below it may be possible to allow for bi-interpretable syntax theories, but surely non-equiconsistent base syntactic theories should be disregarded.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

assumptions). Furthermore, the cases of t-equivalence that I will consider are in fact cases of equality of truth-definitions. One may think of introducing a notion of t-synonymy that relates to t-equivalence like synonymy simpliciter does to bi-interpretability. Given Visser and Friedman's result and the canonical definition of truth translation, however, this would not yield any further insights so I stick with the more general notion of t-equivalence.

I consider again extensions of EA^{T} and extensions of EA with a binary satisfaction predicate. Let $\mathsf{tr}\colon \mathcal{L}_{\mathsf{Sat}} \to \mathcal{L}_{\mathsf{T}}$ and $\mathsf{st}\colon \mathcal{L}_{\mathsf{T}} \to \mathcal{L}_{\mathsf{Sat}}$, where $\mathcal{L}_{\mathsf{Sat}} := \mathcal{L} \cup \{\mathsf{Sat}\}$ and Sat is a binary satisfaction predicate, be truth-translations specified by

$$\begin{aligned} \mathsf{Sat}^{\mathsf{tr}}(x, y) &:\leftrightarrow \mathsf{Tsb}(x, y) \\ \mathsf{T}^{\mathsf{st}}(x) &:\leftrightarrow \mathsf{Sat}(\langle\rangle, x), \end{aligned}$$

where $\langle\rangle$ denotes the empty sequence and the elementary formula $\mathsf{sb}(x, y) = z$ captures the operation of substituting in a formula y its free variables with the result of applying the finite mapping x to the free variables of y. We assume the functions sub and sb to be defined in such a way that in EA we can prove:

$$\mathsf{sub}(\ulcorner\varphi(v)\urcorner, \ulcorner v\urcorner, \overline{x}) = \mathsf{sb}(\langle x\rangle, \ulcorner\varphi(v)\urcorner) \tag{14}$$

$$\forall x\, (\mathsf{sb}(\langle\rangle, x) = x). \tag{15}$$

In investigating positive results on t-retracts and t-equivalence, I start with a simple example to fix the basic reasoning involved and then move to slightly more complex cases of the same sort. Let TBS be EA formulated in $\mathcal{L}_{\mathsf{Sat}}$ plus all instances of the schema

$$\mathsf{Sat}(\langle\rangle, \ulcorner\varphi\urcorner) \leftrightarrow \varphi \tag{tbs}$$

for all $\mathcal{L}$-sentences φ.[29]

PROPOSITION 4.10.

(i) TB *and* TBS *are mutually truth-definable*;

(ii) TB *is a t-retract of* TBS.

Proof. The only thing to check is 4.10(ii). Since we have truth-definitions, we need to take care only of the truth predicate. In TB, $\mathsf{T}^{\mathsf{tr}\circ\mathsf{st}}x$ is equivalent to $\mathsf{Tsb}(\langle\rangle, x)$ that, by (15), is equivalent to $\mathsf{T}x$. The required isomorphism is the identity on TB. □

Here I am not able to employ tr and st to show that TBS is a retract of TB. The reason is that I cannot prove in TBS the expected interaction of substitution and the satisfaction predicate: informally, I cannot prove that the result of substituting the elements of $(x_1, \ldots, x_n)$ in $\varphi(v_1, \ldots, v_n)$ is satisfied by the empty sequence exactly when $\varphi(v_1, \ldots, v_n)$ is satisfied by $(x_1, \ldots, x_n)$. So I let

$$\mathsf{TBS}^* := \mathsf{TBS} \cup \{\forall x, y\, \big(\mathsf{Sat}(\langle\rangle, \mathsf{sb}(x, y)) \leftrightarrow \mathsf{Sat}(x, y)\big)\}.$$

I call D the sentence in curly brackets. That TBS^* is a natural extension of TBS is justified by the following claim:

LEMMA 4.11. TBS^* *and* TB *are mutually truth-definable.*

[29] Here I consider the case in which Δ_0-induction is extended to $\mathcal{L}_{\mathsf{Sat}}$, but the same argument goes through for the corresponding theories with restricted induction.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Proof. That TBS^* defines the truth predicate of TB follows from Proposition 4.10. Also:

$$\mathsf{TB} \vdash \big[\mathsf{Sat}(\langle\rangle, \mathsf{sb}(x, y)) \leftrightarrow \mathsf{Sat}(x, y)\big]^{\mathsf{tr}} \leftrightarrow (\mathsf{Tsb}(\langle\rangle, \mathsf{sb}(x, y)) \leftrightarrow \mathsf{Tsb}(x, y)) \leftrightarrow (\mathsf{Tsb}(x, y) \leftrightarrow \mathsf{Tsb}(x, y)).$$ □

Therefore, we have our first, simple-minded example of t-equivalent theories of truth.

PROPOSITION 4.12.

(i) TBS^* *is a t-retract of* TB;

(ii) TB *and* TBS^* *are t-equivalent.*

Proof. In TBS, $\mathsf{Sat}^{\mathsf{st}\circ\mathsf{tr}}(x, y)$ is provably equivalent to $\mathsf{Sat}(\langle\rangle, \mathsf{sb}(x, y))$. In TBS^* we can then conclude $\mathsf{Sat}(x, y)$. □

The role of D becomes even more prominent if we consider a suitably simplified form of uniform, disquotational satisfaction. UTB has already been introduced on page 18.

DEFINITION 4.13. UTBS *extends* EA^{T} *with all instances of the schema*

$$\forall x \left(\mathsf{Sat}(\langle x\rangle, \ulcorner \varphi(v) \urcorner) \leftrightarrow \varphi(x)\right) \tag{utbs}$$

for all $\mathcal{L}$-formulas φ with the displayed variables free. Similarly as before, we let

$$\mathsf{UTBS}^* := \mathsf{UTBS} + \mathsf{D}.$$

By adapting the reasoning employed for TB and TBS, and by (14) above, we can conclude:

LEMMA 4.14.

(i) UTBS^* *and* UTB *are mutually truth-definable.*

(ii) UTBS^* *and* UTB *are t-equivalent.*

I now consider the case of full Tarskian truth and satisfaction over EA. In EA there are the following elementary formulas expressing the corresponding elementary syntactic notions:

$\mathsf{map}(x) :\Leftrightarrow$ 'x is a finite mapping'

$\mathsf{dom}(x, y) :\Leftrightarrow$ 'y is an element of the domain of the finite mapping x'

$\mathsf{ass}(x, y) :\Leftrightarrow$ 'y is a formula and x is a finite mapping whose domain contains precisely the free variables of y'

$x \sqsupseteq y :\Leftrightarrow$ 'x and y are assignments and $\mathsf{dom}(x)$ contains $\mathsf{dom}(y)$'

DEFINITION 4.15. *The theory* CS *in* $\mathcal{L}_{\mathsf{Sat}}$ *is obtained by extending* EA *formulated in* $\mathcal{L}_{\mathsf{Sat}}$*—that is, in which* Sat *can appear into instances of induction—with the universal closures of the following*:

$$\mathsf{ass}(x, \ulcorner R(v_1, \ldots, v_n)\urcorner) \wedge \Big(\bigwedge_{i=1}^{n} x(\ulcorner v_i \urcorner) = x_i\Big) \rightarrow \big(\mathsf{Sat}(x, \ulcorner R(v_1, \ldots, v_n)\urcorner) \leftrightarrow R(x_1, \ldots, x_n)\big) \tag{S1}$$

$$\mathsf{ass}(x, y) \rightarrow \big(\mathsf{Sat}(x, \mathsf{ng}(y)) \leftrightarrow \neg\, \mathsf{Sat}(x, y)\big) \tag{S2}$$

$$\mathsf{ass}(x, \mathsf{and}(y, z)) \rightarrow \big(\mathsf{Sat}(x, \mathsf{and}(y, z)) \leftrightarrow (\mathsf{Sat}(x, y) \wedge \mathsf{Sat}(x, z))\big) \tag{S3}$$

$$\mathsf{ass}(x, \mathsf{all}(\ulcorner v \urcorner, y)) \rightarrow \big(\mathsf{Sat}(x, \mathsf{all}(\ulcorner v \urcorner, y)) \leftrightarrow (\forall z \sqsupseteq x)\ \mathsf{Sat}(z, y)\big) \tag{S4}$$

$$\mathsf{Sat}(x, y) \rightarrow \mathsf{ass}(x, y). \tag{S5}$$

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

We recall that CT is the theory obtained from CT$\restriction$ from Definition 4.8 by allowing formulas of $\mathcal{L}_{\mathsf{T}}$, in which the truth predicate only applies to formulas of $\mathcal{L}$, into the schema of Δ_0-induction. As one might expect, st and tr give us

LEMMA 4.16.

(i) CT *defines the truth predicate of* CS;

(ii) CS *defines the truth predicate of* CT.

Proof. (i) crucially employs the properties of substitution. For instance, for S1, we reason as follows, with the contextual information that $\mathsf{ass}(x, \ulcorner R(v_1, \ldots, v_n)\urcorner)$ and $\bigwedge_{i=1}^{n} x(\ulcorner v_i \urcorner) = x_i$:

$$\begin{aligned} \mathsf{Tsb}(x, \ulcorner R(v_1, \ldots, v_n)\urcorner) &\leftrightarrow \mathsf{T}\ulcorner R(\dot{x}_1, \ldots, \dot{x}_n)\urcorner && \text{by sb and EA} \\ &\leftrightarrow R(x_1, \ldots, x_n) && \text{by Tat} \end{aligned}$$

For (S4), we notice that EA proves

$$\mathsf{all}(\ulcorner v \urcorner, \mathsf{sub}(x, y)) = \mathsf{sb}(x, \mathsf{all}(\ulcorner v \urcorner, y)). \tag{16}$$

Now if $\mathsf{Tsb}(x, \mathsf{all}(\ulcorner v \urcorner, y))$, also $\mathsf{Tall}(\ulcorner v \urcorner, \mathsf{sb}(x, y))$ and therefore $\forall w\, \mathsf{Tsub}(\mathsf{sb}(x, y), \bar{w})$ by the CT axioms. If there is a mapping $z \supseteq x$ such that $\neg\mathsf{Tsb}(z, y)$, also $\mathsf{ass}(z, y)$ and $\mathsf{ass}(x, \mathsf{all}(\ulcorner v \urcorner, y))$. Therefore, there is a w_0 such that $\neg\, \mathsf{dom}(x, w_0)$ and $\mathsf{dom}(z, w_0)$. Hence, in EA,

$$\mathsf{sub}(\mathsf{sb}(x, y), \bar{w}_0) = \mathsf{sb}(z, y) \tag{17}$$

We can then conclude $\mathsf{Tsb}(z, y)$, contradicting our assumption. The converse direction is obtained similarly, with the help of (16) and (17).

(ii) is obtained once we have established a version of D above via suitable induction on the formal complexity of the 'formula' y:

$$\mathsf{ass}(x, y) \;\rightarrow\; \big(\mathsf{Sat}(x, y) \;\leftrightarrow\; \mathsf{Sat}(\langle\rangle, \mathsf{sb}(x, y))\big) \tag{18}$$

□

PROPOSITION 4.17. CT *and* CS *are truth equivalent.*

Proof. That CT is a retract of CS is immediate by definitions of st and tr. To conclude that CS is a retract of CT, we only need (18). In both cases the required isomorphisms are in fact identities. □

We notice that, unlike the previous observations concerning disquotation, Proposition 4.17 still holds when we formulate CS in a language without domain constants. Moreover, since we essentially employ the extended induction of CS to obtain the t-equivalence of the two theories, it would be interesting to know what happens if we consider induction free versions of the theories involved. The question of whether CS$\restriction$ and CT$\restriction$ are t-equivalent (or mutually truth-definable) is in fact still open.

In desideratum (III) above, we required the notions of t-retracts and t-equivalence to be nonempty. Although not exciting, the examples just presented suffice to accomplish this minimal task. Admittedly, the main role of our proposed sufficient condition for conceptual equivalence is to discourage inadequate claims, and this is mostly accomplished by negative results such as Corollary 4.7.

I now move to an example of *bi-interpretability simpliciter* between truth theories that originates in Leigh & Nicolai (2013), in which theories with external truth were first

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

introduced.[30] Here I only sketch a simplified construction that may be useful to obtain more sophisticated—and more natural—examples. Let EA^2 be obtained by 'cloning' the language $\mathcal{L}$ in a 'two-sorted' version: in practice, I work with relativizing predicates $\mathsf{s}(x)$ and $\mathsf{t}(x)$ that represent the two copies of our numbers. I abbreviate $\forall x(\mathsf{s}(x) \to \varphi(x))$ with $(\forall x : \mathsf{s})\ \varphi(x)$, and similarly for t. The language of EA^2 will thus contain, for instance, two versions of the primitive predicate '… is zero', Z_s and Z_t, and so on for the other primitive notions, including identity symbols $=_s, =_t$. The induction axioms of EA^2 come therefore in two flavours, one in which the relevant variable can only belong to s and one in which it can only belong to t, although formulas from the entire language are allowed in both types of instances. We assume the usual arithmetization of the syntax for EA^2 carried out in the s-portion of the language. We also include in the theory a 'Frege relation' F, that is a relation witnessing an isomorphism between the two domains corresponding to the two sorts. In other words we add axioms of the form:

$$x\mathsf{F}y \ \to\ \mathsf{s}(x) \wedge \mathsf{t}(y) \tag{F0}$$

'F is a bijection between s-objects and t-objects' (F1)

$$(\forall x : \mathsf{s})(\forall y : \mathsf{t})\big(x\ \mathsf{F}\ y\ \to\ (\ \mathsf{Z}_\mathsf{s}(x) \leftrightarrow \mathsf{Z}_\mathsf{t}(y))\big) \tag{F2}$$

$$(\forall x, u : \mathsf{s})(\forall y, v : \mathsf{t})\big(x\ \mathsf{F}\ y\ \wedge u\ \mathsf{F}\ v\ \to\ (\mathsf{S}_\mathsf{s}(x, u) \leftrightarrow \mathsf{S}_\mathsf{t}(y, v))\big) \tag{F3}$$

By formal induction in EA^2, the analogues of F2 and F3 for the predicates A and M for addition, multiplication and exponentiation come out as theorems of EA^2:

$$(\forall x, y, z : \mathsf{s})(\forall u, v, w : \mathsf{t})(x\ \mathsf{F}\ u\ \wedge y\ \mathsf{F}\ v\ \wedge z\ \mathsf{F}\ w\ \to\ (\mathsf{A}_\mathsf{s}(x, y, z) \leftrightarrow \mathsf{A}_\mathsf{t}(u, v, w)) \tag{19}$$

$$(\forall x, y, z : \mathsf{s})(\forall u, v, w : \mathsf{t})(x\ \mathsf{F}\ u\ \wedge y\ \mathsf{F}\ v\ \wedge z\ \mathsf{F}\ w\ \to\ (\mathsf{M}_\mathsf{s}(x, y, z) \leftrightarrow \mathsf{M}_\mathsf{t}(u, v, w)) \tag{20}$$

$$(\forall x, y : \mathsf{s})(\forall u, v : \mathsf{t})(x\ \mathsf{F}\ u\ \wedge y\ \mathsf{F}\ v\ \to\ (\mathsf{E}_\mathsf{s}(x, y) \leftrightarrow \mathsf{E}_\mathsf{t}(u, v))). \tag{21}$$

I then introduce *typed* truth axioms for EA^2 by employing a truth predicate Tr of type s, that is, applying only to objects of sort s. The characterizing truth axioms, besides the obvious analogues of T2 and $(\star)$ (cf. p. 18), are the ones for atomic formulas and quantifiers:

$$\mathsf{Tr}{\ulcorner R_\mathsf{s}(\dot{x}_1, \ldots, \dot{x}_n)\urcorner} \leftrightarrow R_\mathsf{t}\big(\mathsf{F}(x_1), \ldots, \mathsf{F}(x_n))\big) \ \text{ for all } \mathsf{s}\text{- and } \mathsf{t}\text{-relation symbols} \tag{Tat*}$$

$$\mathsf{Fml}^1_{\mathcal{L}^\mathsf{s}}(y) \to \Big(\mathsf{Tr}\,(\mathsf{all}_\mathsf{s}({\ulcorner v \urcorner}, y)) \leftrightarrow (\forall z : \mathsf{s})\,\big(\mathsf{Tr}\ \ \mathsf{sub}_\mathsf{s}(y, \overline{z})\big)\Big). \tag{Tq}$$

I refer to the resulting theory of truth as $\mathsf{T}[\mathsf{EA}^2]$. The theory CT has been already introduced: it is convenient to think of the current version of CT as built over a one-sorted language of sort s, a sublanguage of the language of EA^2.

PROPOSITION 4.18. $\mathsf{T}[\mathsf{EA}^2]$ *and* CT *are bi-interpretable.*

[30] Although already at that time the influential work of Richard Heck, who should be granted the priority, together with Albert Visser, of introducing external truth, was widely circulating. This approach can be motivated in the following way: If one starts with an arbitrary object theory U, unlike what it is commonly thought, it is not so straighforward to introduce a notion of truth for U. If, for instance, U fails to be sequential (cf. §1), the theory does not have a good notion of sequence and we won't have the necessary resources to express full satisfaction in a direct or derivative form—e.g., by employing domain constants as in the case of a unary truth predicate. However, truth and sequences can be added 'from the outside' (Leigh & Nicolai, 2013; Heck, 2015; Nicolai, 2016a).

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Proof. I specify the interpretation $\mathsf{K}\colon \mathsf{T}[\mathsf{EA}^2] \to \mathsf{CT}$. The idea behind it is entirely straightforward: the two copies of the numbers in $\mathsf{T}[\mathsf{EA}^2]$ are reproduced in CT as pairs with 0 and 1 as first members, whereas primitive relations collapse into their counterparts in CT. In CT we find the following elementary formulas:

$$\mathsf{pair}(x) :\leftrightarrow \text{‘}x \text{ is an ordered pair’}$$
$$\mathsf{pr}i(x) :\leftrightarrow \text{‘}x \text{ is an ordered pair with first member } i\text{’}$$
$$\pi_i(x, y) :\leftrightarrow \text{‘}y \text{ is the } i^{th} \text{ projection of the pair } x\text{’}$$

I employ $\pi_i(\cdot)$ in a functional form. Let

$$\mathsf{t}^{\mathsf{K}}(x) :\leftrightarrow x : \mathsf{pr0} \qquad \mathsf{s}^{\mathsf{K}}(x) :\leftrightarrow x : \mathsf{pr1}$$
$$\mathsf{Z}_{\mathsf{t}}^{\mathsf{K}}(x) :\leftrightarrow \mathsf{Z}(\pi_1(x)) \qquad \mathsf{Z}_{\mathsf{s}}^{\mathsf{K}}(x) :\leftrightarrow \mathsf{Z}(\pi_1(x))$$
$$\vdots \qquad\qquad \vdots$$
$$\mathsf{Tr}^{\mathsf{K}}x :\leftrightarrow \mathsf{T}\pi_1(x) \qquad \mathsf{F}^{\mathsf{K}}(x, y) :\leftrightarrow x : \mathsf{pr1} \wedge y : \mathsf{pr0} \wedge \pi_1(x) = \pi_1(y)$$
$$x =_{\mathsf{s}} y :\leftrightarrow x, y : \mathsf{pr1} \wedge \pi_1(x) = \pi_1(y) \qquad x =_{\mathsf{t}} y :\leftrightarrow x, y : \mathsf{pr0} \wedge \pi_1(x) = \pi_1(y)$$

K relativizes quantifiers to $\mathsf{pair}(x)$. The vertical dots refer to the clauses for the other arithmetical relations (in both versions), and the convention employed in the definition of EA^2 concerning relativized quantifiers has been extended to $\mathsf{pr0}$ and $\mathsf{pr1}$.

The interpretation $\mathsf{L}\colon \mathsf{CT} \to \mathsf{T}[\mathsf{EA}^2]$ simply relativizes everything to the predicate s. Crucially:

$$(\forall x \varphi)^{\mathsf{L}} :\leftrightarrow (\forall x : \mathsf{s})\varphi^{\mathsf{L}}$$
$$\mathsf{T}^{\mathsf{L}}(x) :\leftrightarrow \mathsf{Tr}(x)$$

By letting, in $\mathsf{T}[\mathsf{EA}^2]$,

$$\mathsf{G}(x, y) :\leftrightarrow (\mathsf{pr1}^{\mathsf{s}}(x) \wedge \pi_1^{\mathsf{s}}(x) = y) \vee (\mathsf{pr0}^{\mathsf{s}}(x) \wedge \mathsf{F}(\pi_1^{\mathsf{s}}(x)) = y),$$

it is not difficult to check that $\mathsf{G}\colon \mathsf{L} \circ \mathsf{K} \cong \mathsf{id}_{\mathsf{T}[\mathsf{EA}^2]}$ and $\pi_1\colon \mathsf{K} \circ \mathsf{L} \cong \mathsf{id}_{\mathsf{CT}}$, witnessing the required bi-interpretability of $\mathsf{T}[\mathsf{EA}^2]$ and CT. □

We conclude this section by with a chart summarizing our findings.

DESIDERATUM	EXAMPLES
Extensions of mutual truth-definability	t-retracts, t-equivalence
Separating t-retracts, t-equivalence, mutual truth-definability	PUTB, KF; PUTB↾, KF↾
Non-emptiness of t-retracts, t-equivalence	Flavours of Truth and Satisfaction
Bi-interpretability simpliciter	‘External’ truth

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

§5. Extending the framework: syntactical embeddings, e-retractions, e-equivalence. As anticipated in the initial section, natural generalizations of the notions of t-retract and t-equivalence may yield new insights on the comparison between the operations of adding typed truth and predicative comprehension to a ground syntactic structure. Nicolai (2016b) focuses on abstracting away restrictive choices of the object theory. I now consider a less general framework but stricter notions of reduction. In particular, claims like the following belong to the truth-theoretic and foundational folklore:

> Typing truth predicates corresponds to a much more severe move in the case of comprehension: typing corresponds to predicative typed comprehension [...] Actually ramified type theory over Peano arithmetic as base theory, which is known as ramified analysis, is *equivalent* to typed compositional truth. (Halbach, 2014, p. 28)

In what sense should this 'equivalence' be understood? Most likely Halbach refers to *proof-theoretical equivalence* (cf. Feferman (1988)), which in the cases I consider below is no different from a more liberal notion of truth-definition. In this section I employ analogues of the notions of truth-definitions, t-retractions and t-equivalence to refine these folklore statements. Several observations contained in this section rely on unpublished work of Ali Enayat and Albert Visser.

I first generalize the notion of a truth-definition to arbitrary nonlogical vocabulary extending the language of a suitable syntax theory B.

DEFINITION 5.1 (Syntactical embedding). *Let $\mathcal{L}_B \subseteq \mathcal{L}_T, \mathcal{L}_W$ and $B \subseteq T, W$. We say that T is syntactically embeddable in W if and only if there is a relative interpretation $K : T \to W$ that leaves the primitive vocabulary of $\mathcal{L}_B$ unchanged and does not (non-trivially) relativize its quantifiers.*

Truth-definitions as defined above are clearly examples of syntactical embeddings. We now move to generalizations of the notions of t-retract and t-equivalence.

DEFINITION 5.2 (e-retract, e-equivalence). *Let $T, W \supseteq B$ be given.*

(i) *T is an e-retract of W if there are syntactical embeddings $K : T \to W$ and $L : W \to T$ and a T-definable $F : L \circ K \cong \mathsf{id}_T$;*
(ii) *T and W are e-equivalent if there are syntactical embeddings $K : T \to W$ and $L : W \to T$, a T-definable $F : L \circ K \cong \mathsf{id}_T$ and a W-definable $G : K \circ L \cong \mathsf{id}_W$.*

To be able to apply the new notions, I briefly recall some standard definitions concerning subsystems of second-order arithmetic as they can be found, for instance, in Simpson (2009). $\mathcal{L}_2$ is the two-sorted language extending $\mathcal{L}$ with a sort for sets of natural numbers, or 'reals'. The two kinds of variables will often be denoted by $x, y, z, \ldots$ and $X, Y, Z \ldots$ respectively, possibly with indices. In practice, to conform with our notions of reduction, it is convenient to consider the two-sorted language as notational abbreviation for a language with suitable relativizing predicates. That is, we officially work in a single-sorted language with relativizing predicates $\mathsf{se}(x)$, 'x is a set', and $\mathsf{nu}(x)$, 'x is a number', and a primitive membership predicate $\in$. A formula of $\mathcal{L}_2$ is said to be *arithmetical* if it contains no set quantifiers but possibly set parameters. The so-called *arithmetical comprehension schema* has the form

$$\exists X \, \forall x \, (x \in X \leftrightarrow \varphi(x, \vec{u}, \vec{Y})), \tag{aca}$$

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

where $\varphi(v_0)$ is arithmetical and does not contain X.[31] The theory ACA in $\mathcal{L}_2$ is obtained by relativizing the basic axioms of EA to the 'numbers' sort and by extending it with (aca) and an induction schema *for arbitrary* $\mathcal{L}_2$*-formulas*. $\mathsf{ACA}^{\mathsf{pf}}$ in $\mathcal{L}_2$ is obtained from ACA by disallowing set-parameters into instances of the comprehension schema. The theory ACA_0 is obtained from ACA by replacing its full induction schema with the equivalent, in the official language, of the single sentence

$$(0 \in X \wedge \forall x\,(x \in X \rightarrow \mathsf{S}x \in X)) \rightarrow \forall x\; x \in X. \qquad \mathsf{IND}$$

In the official presentation of the theories by means of relativizing predicates, I also require the domain to be partitioned by them.

For the purpose of this section I will consider only *finite* iterations of full predicative comprehension obtained intuitively by iterating ACA n-times. The restriction to finite levels is motivated by the difficulties in the formulation of iterations of predicative comprehension for higher ordinals. In particular, ramified analysis is classically formulated by employing reflection principles (see Feferman (1964)), and this creates some difficulties in transferring the results below to the transfinite.

More formally, we have

$$\mathcal{L}^{<0} := \mathcal{L}$$
$$\mathcal{L}^{<n+1} := \mathcal{L} \cup \{\mathsf{se}^i, \in^i\} \qquad \text{for } i \leq n$$

I will also write $\mathcal{L}^n$ for $\mathcal{L}^{<n+1}$. RA_0 is ACA itself. The *degree* of a formula of $\mathcal{L}^n$ is the maximum of the $k+1$ such that $\mathsf{se}^k(y)$ appears in the formula with y bound and of the m such that $\mathsf{se}^m(z)$ appears in the formula with z is free (of course with $m \leq n$ and $k < n$). I write φ^n for a formula of at most degree n. Then RA_{n+1} results from RA_n by extending its induction schema to the new language and by adding to it the axioms:

$$x \in^{n+1} y \rightarrow \mathsf{se}^{n+1}(y) \tag{22}$$
$$\exists y\,(\mathsf{se}^{n+1}(y) \wedge \forall x(x \in^{n+1} y \leftrightarrow \varphi^{n+1}(x, \vec{u}, \vec{X}^i))). \tag{23}$$

For reasons of readability, we have omitted the relativization $\mathsf{nu}(x)$. We will keep this convention in what follows. In (23), y is not free in φ and $\vec{X}^i$ is a string of parameters from elements of se^i for $i \leq n+1$. By disallowing (set-)parameters into the schema (23), one obtains a parallel hierarchy of finitely iterated (parameter-free) predicative comprehension, with $\mathsf{RA}_0^{\mathsf{pf}} := \mathsf{ACA}^{\mathsf{pf}}$.

We are interested in associating the theories just presented with finite iterations of $\mathsf{UTB}[\mathsf{PA}]$ and $\mathsf{CT}[\mathsf{PA}]$. The latter theories are formulated in the languages $\mathcal{L}_{\mathsf{T}}^n := \mathcal{L} \cup \{\mathsf{T}_0, \ldots, \mathsf{T}_n\}$ for $n \in \omega$, with $\mathcal{L}_{\mathsf{T}}^{\leq 0} := \mathcal{L}$ and $\mathcal{L}_{\mathsf{T}}^{<n+1} := \mathcal{L}_{\mathsf{T}}^n$.

DEFINITION 5.3. *Let* $\mathsf{RT}_0 := \mathsf{CT}[\mathsf{PA}]$ *and* $\mathsf{RDT}_0 := \mathsf{UTB}[\mathsf{PA}]$*, both formulated in* $\mathcal{L}_{\mathsf{T}}^0$.

- *The theory* RDT_{n+1} *is obtained by extending* RDT_n *with full* $\mathcal{L}_{\mathsf{T}}^{n+1}$*-induction and all instances of*

$$\forall x\,\big(\mathsf{T}_{n+1}{}^{\ulcorner}\varphi(\dot{x})^{\urcorner} \leftrightarrow \varphi(x)\big)$$

for all $\mathcal{L}_n$*-formulas* $\varphi(v)$.

[31] In our official formulation, the arithmetical comprehension thus becomes

$$\mathsf{se}(\vec{z}) \rightarrow \exists y\big(\mathsf{se}(y) \wedge \forall u\big(\mathsf{nu}(u) \rightarrow (u \in y \leftrightarrow \varphi(u, \vec{v}, \vec{z}))\big)\big).$$

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

- *The theory* RT_{n+1} *contains the axioms of* RT_n*, including the induction schema for the entire language* $\mathcal{L}_{\mathsf{T}}^{n+1}$*, and the following, for* $m < n+1$:

$$\forall x_1, \dots, x_n \left(\mathsf{T}_{n+1}{}^{\ulcorner} R(\dot{x}_1, \dots, \dot{x}_n)^{\urcorner} \leftrightarrow R(x_1, \dots, x_n)\right) \tag{$\mathsf{R1}_{n+1}$}$$

$$\forall x \left(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{<m}}(x) \rightarrow (\mathsf{T}_{n+1}{}^{\ulcorner}\mathsf{T}_m\, \dot{x}^{\urcorner} \leftrightarrow \mathsf{T}_m\, x)\right) \tag{$\mathsf{R2}_{n+1}$}$$

$$\forall x \left(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{n}}(x) \rightarrow (\mathsf{T}_{n+1}\mathsf{ng}(x) \leftrightarrow \neg \mathsf{T}_{n+1} x)\right) \tag{$\mathsf{R3}_{n+1}$}$$

$$\forall x, y \left(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{n}}(\mathsf{and}(x, y)) \rightarrow (\mathsf{T}_{n+1}\mathsf{and}(x, y) \leftrightarrow (\mathsf{T}_{n+1} x \wedge \mathsf{T}_{n+1} y))\right) \tag{$\mathsf{R4}_{n+1}$}$$

$$\forall x, v \left(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{n}}(\mathsf{all}(v, x)) \rightarrow (\mathsf{T}_{n+1}\mathsf{all}(v, x) \leftrightarrow \forall y\, \mathsf{T}_{n+1}\mathsf{sub}(x, \overline{y}))\right) \tag{$\mathsf{R5}_{n+1}$}$$

$$\forall y \leq n\, \forall x \left(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{<\dot{y}}}(x) \rightarrow (\mathsf{T}_{n+1}{}^{\ulcorner}\mathsf{T}_{\dot{y}}\, \dot{x}^{\urcorner} \leftrightarrow \mathsf{T}_{n+1}\, x)\right). \tag{$\mathsf{R6}_{n+1}$}$$

The standard reductions between ramified truth and predicative comprehension mainly amount to mutual syntactical embeddings (cf. Halbach (2014); Feferman (1991); Takeuti (1987); Fischer (2009)): Let $\mathbb{T}_0(X^0, u)$ be the $\mathcal{L}_0$-formula

$$\begin{aligned}
&\forall y \big(y \in X^0 \leftrightarrow \mathsf{Sent}_{\mathcal{L}}(y) \wedge \mathsf{lc}(x) \leq u \wedge \\
&\forall \vec{x}\, (y = {}^{\ulcorner} R(\dot{\vec{x}})^{\urcorner} \rightarrow (y \in X^0 \leftrightarrow R\vec{x})) \\
&\forall z\, (y = \mathsf{ng}(z) \rightarrow (y \in X^0 \leftrightarrow z \notin X^0)) \wedge \\
&\forall w, z\, (y = \mathsf{and}(w, z) \rightarrow (y \in X^0 \leftrightarrow (w \in X^0 \wedge z \in X^0))) \wedge \\
&\forall v, z, y\, (x = \mathsf{all}(v, z) \rightarrow (x \in X^0 \leftrightarrow \forall y\, (\mathsf{sub}(z, \overline{y}) \in X^0)))\big).
\end{aligned}$$

Moreover, let $\mathbb{V}_0(x) :\leftrightarrow (\exists Y^0)(\mathbb{T}_0(Y^0, \mathsf{lc}(x)) \wedge x \in Y^0)$. This latter formula plays the role of a partial truth predicate for $\mathcal{L}$-formulas of complexity up to (and including) the complexity of x. To define what it means to be a partial truth set of level $n+1$, we generalize the notion of logical complexity of a formula to sentences of $\mathcal{L}_{\mathsf{T}}^{n}$ by letting $\mathsf{lc}^{n+1}({}^{\ulcorner}\mathsf{T}_n\, \dot{y}^{\urcorner})$ to be 0 and keeping the rest as it is:

$$\begin{aligned}
\mathbb{T}_{n+1}(X^{n+1}, u) :\leftrightarrow\ & \forall y \big(y \in X^{n+1} \leftrightarrow \mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{n}}(y) \wedge \mathsf{lc}^{n+1}(y) \leq u \wedge \\
& \forall \vec{x}\, (y = {}^{\ulcorner} R(\dot{\vec{x}})^{\urcorner} \rightarrow (y \in X^{n+1} \leftrightarrow R\vec{x})) \wedge \\
& \forall z\, (y = \mathsf{ng}(z) \rightarrow (y \in X^{n+1} \leftrightarrow z \notin X^{n+1})) \wedge \\
& \forall w, z\, (y = \mathsf{and}(w, z) \rightarrow (y \in X^{n+1} \leftrightarrow (w \in X^{n+1} \wedge z \in X^{n+1}))) \wedge \\
& \forall v, z, y\, (x = \mathsf{all}(v, z) \rightarrow (x \in X^{n+1} \leftrightarrow \forall y\, (\mathsf{sub}(z, \overline{y}) \in X^{n+1}))) \wedge \\
& \forall z\, (\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{<n}}(z) \wedge y = {}^{\ulcorner}\mathsf{T}_n \dot{z}^{\urcorner} \rightarrow (y \in X^{n+1} \leftrightarrow \mathbb{V}_n(z))) \wedge \\
& \forall u \leq n\, \forall z\, (\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{<\dot{u}}}(z) \wedge y = {}^{\ulcorner}\mathsf{T}_{\dot{u}} \dot{z}^{\urcorner} \rightarrow (y \in X^{n+1} \leftrightarrow z \in X^{n+1}))\big).
\end{aligned}$$

The formula $\mathbb{V}_{n+1}(x)$ is then defined in the obvious way.

We finally adapt the folklore translations to our setting by defining the translation functions $\mathsf{K}_n \colon \mathcal{L}_{\mathsf{T}}^{n} \rightarrow \mathcal{L}^n$, $\mathsf{L}_n \colon \mathcal{L}^n \rightarrow \mathcal{L}_{\mathsf{T}}^{n}$ in Table 1. We remark that $\mathcal{L}^n$ and $\mathcal{L}_{\mathsf{T}}^{n}$ feature only one identity symbol.

In the table, the elementary formula $\mathsf{sub}^1(x, y) = z$ expresses the result of formally replacing the single free variable of a formula y with x. The folkore reductions can be summarized as follows.

PROPOSITION 5.4.

(i) *For each* $n \in \omega$, RDT_n *and* $\mathsf{RA}_n^{\mathsf{pf}}$ *are mutually syntactically embeddable.*

(ii) *For each* $n \in \omega$, RT_n *and* RA_n *are mutually syntactically embeddable.*

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Table 1. *The translations* K_n, L_n

$R^{\mathsf{K}_n}(\vec{x}) :\leftrightarrow R\vec{x}$	$R^{\mathsf{L}_n}(\vec{x}) :\leftrightarrow R\vec{x}$	for $R \in \mathcal{L}$
$x =^{\mathsf{K}_n} y :\leftrightarrow x = y$	$x =^{\mathsf{L}_n} y :\leftrightarrow x = y$	
$\mathsf{T}_m^{\mathsf{K}_n} x :\leftrightarrow \mathbb{V}_m(x)$	$x(\in^m)^{\mathsf{L}_n} y :\leftrightarrow \mathsf{T}_m\mathsf{sub}^1(x, y)$	with $m \leq n$
$(\neg\varphi)^{\mathsf{K}_n} :\leftrightarrow \neg\varphi^{\mathsf{K}_n}$	$(\neg\varphi)^{\mathsf{L}_n} :\leftrightarrow \neg\varphi^{\mathsf{L}_n}$	
$(\varphi \wedge \psi)^{\mathsf{K}_n} :\leftrightarrow \varphi^{\mathsf{K}_n} \wedge \psi^{\mathsf{K}_n}$	$(\varphi \wedge \psi)^{\mathsf{L}_n} :\leftrightarrow \varphi^{\mathsf{L}_n} \wedge \psi^{\mathsf{L}_n}$	
$(\forall x\varphi)^{\mathsf{K}_n} :\leftrightarrow (\forall x : \mathsf{nu})\,\varphi^{\mathsf{K}_n}$	$(\forall x\varphi)^{\mathsf{L}_n} :\leftrightarrow \forall x\,\varphi^{\mathsf{L}_n}$	
	$\mathsf{nu}^{\mathsf{L}_n}(x) :\leftrightarrow x = x$	
	$(\mathsf{se}^m)^{\mathsf{L}_n}(x) :\leftrightarrow \mathsf{Fml}^1_{\mathcal{L}^{<m}}(x)$	$m \leq n$

Proof. One can check that K_n and L_n are the required syntactical embeddings, given the following facts:

$$\mathsf{RA}_n^{\mathsf{pf}} \vdash \exists X^m\, \mathbb{T}_m(X^m, \overline{k}) \qquad \text{for } m \leq n \text{ and } k \in \omega \tag{24}$$

$$\mathsf{RA}_n \vdash \forall x\, \exists X^n\, \mathbb{T}_n(X^n, x) \tag{25}$$

$$\text{if } \varphi(v) \text{ is in } \mathcal{L}^m, \text{ then } \varphi^{\mathsf{L}_m}(v) \text{ belongs to } \mathcal{L}^m_{\mathsf{T}} \tag{26}$$

$$\mathsf{RT}_n \vdash \forall u, \vec{z}\,(\mathsf{T}_m{}^{\ulcorner}\varphi^{\mathsf{L}_n}(\dot{u}, \dot{\vec{z}})^{\urcorner} \leftrightarrow \varphi^{\mathsf{L}_n}(u, \vec{z})) \qquad \text{for } m \leq n, \varphi \text{ in } \mathcal{L}^{<m}, \text{ and } \mathsf{se}^{<m}(\vec{z}) \tag{27}$$

□

From the previous chapter we know that e-retractions and e-equivalences properly strengthen mutual syntactical embeddings. It is thus interesting to investigate whether the folklore mutual embeddings between typed truth and comprehension can be lifted to stricter notions of equivalence. The following shows that the folklore translations partially suffice.

PROPOSITION 5.5. *For each* $n \in \omega$, RT_n *is an e-retract of* RA_n.

Proof. The case $n = 0$ follows from the fact that in CT[PA] we have partial truth predicates $\mathsf{Tr}_{\Sigma_n}(\cdot)$ of complexity Σ_n for Σ_n-formulas and for each n. (Hájek & Pudlák, 1998, I.(d)). By its extended induction, we can prove in CT[PA] that

$$\mathsf{T}^{\mathsf{L}_0 \circ \mathsf{K}_0}(x) \leftrightarrow \exists y\, \mathsf{T}_0{}^{\ulcorner}\mathsf{Tr}_{\Sigma_{\dot{y}}}(\dot{x})^{\urcorner} \leftrightarrow \mathsf{T}_0 x. \tag{28}$$

We emphasize that in (28) the truth predicate T_0 enables us to quantify over the index of the partial truth predicate. The required isomorphism is thus the identity on RT_0.

The same idea can be extended to the languages $\mathcal{L}_n$, by focusing now on the hierarchy of $\Sigma_m^{\mathsf{T}_n}$-formulas, and constructing in the standard way partial truth predicates $\mathsf{Tr}_{\Sigma_v^{\mathsf{T}_n}}(x)$ for the languages $\mathcal{L}^n_{\mathsf{T}}$; they should be thought as formalizing the stages of the construction of T_{n+1} from previsouly defined truth predicates. In particular, all $\mathcal{L}^n_{\mathsf{T}}$-sentences φ deemed true at previous stages are such that $\mathsf{T}_{n+1}{}^{\ulcorner}\mathsf{Tr}_{\Sigma_{\dot{0}}^{\mathsf{T}_n}}(\ulcorner\varphi\urcorner)^{\urcorner}$ is provable in RT_{n+1}.

Again by full $\mathcal{L}^{n+1}_{\mathsf{T}}$-induction:

$$\mathsf{RT}_{n+1} \vdash \mathsf{T}^{\mathsf{L}_{n+1} \circ \mathsf{K}_{n+1}}(x) \leftrightarrow \exists y\, \mathsf{T}_{n+1}{}^{\ulcorner}\mathsf{Tr}_{\Sigma_{\dot{y}}^{\mathsf{T}_n}}(\dot{x})^{\urcorner} \leftrightarrow \mathsf{T}_{n+1}(x). \tag{29}$$

As before, the required isomorphism is indeed the identity function on RT_{n+1}. □

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

But Proposition 5.5 is in a sense the best we can do. I adapt to the present setting an unpublished argument by Enayat and Visser and show that ramified analysis, both in full or parameter-free form, cannot be a retract of ramified truth, both in full compositional or uniform disquotational form.[32]

PROPOSITION 5.6.

(i) $\mathsf{RA}_n^{\mathsf{pf}}$ *is not a retract nor an e-retract of* $\mathsf{RDT}_n[\mathsf{PA}]$*, for* $n \in \omega$.

(ii) RA_n *is not a retract nor an e-retract of* $\mathsf{RT}_n[\mathsf{PA}]$.

Proof. I give the argument for 5.6(i) the same reasoning applies to 5.6(ii). If $\mathsf{RA}_n^{\mathsf{pf}}$ were a retract of RDT_n, *any* model $(\mathcal{M}_0, \mathcal{R}_0, \ldots, \mathcal{R}_n)$ of $\mathsf{RA}_n^{\mathsf{pf}}$—where the $\mathcal{R}_i$'s are subsets of $\mathcal{M}_0$—could define an internal model $(\mathcal{M}_1, \mathcal{S}_0, \ldots, \mathcal{S}_n) \vDash \mathsf{RDT}_n$ of $\mathcal{M}_0$ that would in turn define within itself an internal model $(\mathcal{M}_2, \mathcal{T}_0, \ldots, \mathcal{T}_n) \vDash \mathsf{RA}_n^{\mathsf{pf}}$ of $\mathcal{M}_1$ such that $\mathcal{M}_0$ is isomorphic to $\mathcal{M}_2$, definably in $\mathcal{M}_0$ (cf. Figure 1, §3).

Let us now start with the 'standard' model of $\mathsf{RA}_n^{\mathsf{pf}}$, i.e., the tuple $(\omega, \mathcal{P}(\vec{\omega}))$. By the above considerations, we have the situation depicted in Figure 2, where $(\mathcal{M}_1, \vec{\mathcal{S}}) \vDash \mathsf{RDT}_n$ and $(\mathcal{M}_2, \vec{\mathcal{T}}) \vDash \mathsf{RA}_n^{\mathsf{pf}}$—obviously the $\mathcal{S}_i$ are subsets of $\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}^{<n}}$ and the $\mathcal{T}_i$'s subsets of $\mathcal{M}_2$. Since $(\mathcal{M}_2, \vec{\mathcal{T}}) \cong (\omega, \mathcal{P}(\vec{\omega}))$, there is an interpretation of $(\omega, \mathcal{P}(\omega))$ in $(\mathcal{M}_1, \mathcal{S})$. Thus $(\mathcal{M}_1, \mathcal{S})$ defines its standard natural numbers. At the same time, since $(\mathcal{M}_1, \vec{\mathcal{S}})$ satisfies induction *with the truth predicate* and it interprets the basic axioms of EA, by Lemma 2.2 one can $(\mathcal{M}_1, \vec{\mathcal{S}})$-define an injection $\mathsf{f}\colon \mathcal{M}_1 \to \omega$ by primitive recursion. Therefore $\mathcal{M}_1$ is countable, and thus it cannot define the uncountable structure $(\omega, \mathcal{P}(\vec{\omega}))$. □

COROLLARY 5.7.

(i) *For each* $n \in \omega$, $\mathsf{RA}_n^{\mathsf{pf}}$ *and* RDT_n *are not e-equivalent nor bi-interpretable*;

(ii) *For each* $n \in \omega$, RA_n *and* RT_n *are not e-equivalent nor bi-interpretable*;

COROLLARY 5.8. *The relations 'being an e-retract of' and e-equivalence are properly stricter than mutual syntactical embeddability.*

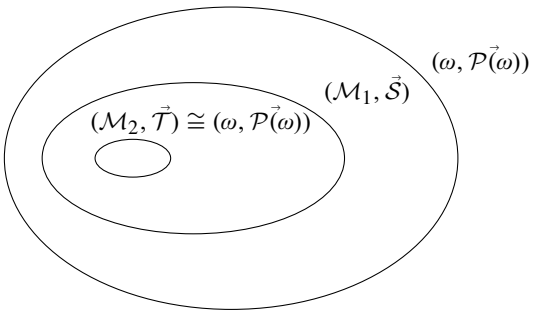


Fig. 2. Models in the proof of Proposition 5.6(i).

[32] Note that the argument is so general that can be applied also to the transfinite, once a suitable formulation of ramified analysis is given. I thank Albert Visser for giving me the permission of quoting his unpublished work.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Corollary 5.7 is already worth consideration. It denies the possibility of strong forms of equivalence between predicative comprehension and typed truth. Typed compositional truth predicates seem to be 'stronger', in the precise sense described above, than the corresponding set-theoretic axioms. This gives us already a refinement of the folklore claim of the correspondence between typed truth and predicative comprehension, at least in the case of full compositional truth. A natural question therefore is to investigate the nature of this asymmetry by measuring what exactly one has to add to ramified analysis to recapture the corresponding truth predicates in the sense of e-equivalence. To do this, at least in the case of RA_n and RT_n, I again elaborate on an idea by Enayat and Visser.

I move to definitional extensions of the RA_n's obtained by adding identity predicates $=_m$, for $m \leq n$, satisfying

$$y =_m z \leftrightarrow \forall u(u \in^m y \leftrightarrow u \in^m z) \tag{30}$$

$$y =_m z \rightarrow \mathsf{se}^m(y) \wedge \mathsf{se}^m(z). \tag{31}$$

For simplicity, I again employ RA_n and $\mathsf{RA}_n^{\mathsf{pf}}$ as names of the extended theories. Also I keep K_m from Table 1 fixed. In RT_n, for $m \leq n$, I define

$$x \sim_m y :\leftrightarrow \mathsf{Fml}^1_{\mathcal{L}_{\mathsf{T}}^{\leq m}}(x) \wedge \mathsf{Fml}^1_{\mathcal{L}_{\mathsf{T}}^{\leq m}}(y) \wedge \forall u\,(\mathsf{T}_m\mathsf{sub}^1(u, x) \leftrightarrow \mathsf{T}_m\mathsf{sub}^1(u, y)) \tag{32}$$

(32) simply says that the $\mathcal{L}_{\mathsf{T}}^{\leq m}$-formulas x, y are satisfied by the same objects. The translation M_m is obtained from L_m in Table 1 by introducing clauses for $=_m$ as follows, with $k \leq m$:

$$x =_k^{\mathsf{M}_m} y :\leftrightarrow x \sim_k y. \tag{33}$$

M_n determines a syntactical embedding of RA_n and $\mathsf{RA}^{\mathsf{pf}}$ in RA_n and RDT_n respectively. Moreover, the argument in Proposition 5.5 proceeds unchanged if M_n is employed instead of L_n. Let

$$\mathsf{D}_m :\leftrightarrow \bigwedge_{i=0}^{m} \forall x \left(\mathsf{se}^i(x) \rightarrow \exists y \left(\mathsf{Fml}_{\mathcal{L}_T^{<i}}(y) \wedge \forall u(u \in^i x \leftrightarrow \mathbb{V}_i(\mathsf{sub}^1(u, y)))\right)\right).$$

The sentence D_n says that for all $i \leq n$, every set of level i can be defined by a $\mathcal{L}_{\mathsf{T}}^{<i}$-formula: by 'define' here I mean that its elements can be shown to be exactly the elements satisfying this formula. By reflecting on the proof of Proposition 5.6, one can immediately see how the addition of D_n excludes what we called the 'standard model of RA_n' from the space of the models of $\mathsf{RA}_n + \mathsf{D}_n$. Moreover, D_n forces the $\mathcal{L}_{\mathsf{T}}^{<i}$-formula defining a set X^i to be unique up to $\sim_i^{\mathsf{K}_i}$-equivalence. By employing essentially the same reasoning as in Proposition 5.5, we have

LEMMA 5.9. *For each $n \in \omega$,*

(i) M_n *is a syntactical embedding of* $\mathsf{RA}_n + \mathsf{D}_n$ *in* RT_n;

(ii) RT_n *is an e-retract of* $\mathsf{RA}_n + \mathsf{D}_n$.

Finally we can check that the sentences D_n suffice to restore the symmetry between the two hierarchies:

PROPOSITION 5.10. $\mathsf{RA}_n + \mathsf{D}_n$ *is an e-retract of* RA_n.

Proof. We know that $\mathsf{RA}_n + \mathsf{D}_n$ and RT_n are mutually syntactically embeddable via K_n and M_n. The induction schema of $\mathsf{RA}_n + \mathsf{D}_n$ enables one to prove, for all $m \leq n$ (and for all $n \in \omega$),

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

$$\forall y\Big(\mathsf{Fml}_{\mathcal{L}_{\mathsf{T}}^{<m}}(y) \to (\exists X^m)(\forall u)(u \in^m X^m \leftrightarrow \mathbb{V}_m(\mathsf{sb}(u,y)))\Big). \tag{34}$$

Next I let

$$\mathsf{H}_n(x,y) :\leftrightarrow \bigvee_{i=0}^{n}(\mathsf{Fml}_{\mathcal{L}_{\mathsf{T}}^{<i}}(x) \wedge \mathsf{se}^i(y) \wedge (\forall u)(u \in^i y \leftrightarrow \mathbb{V}_i(\mathsf{sub}^1(u,x)))) \vee \tag{35}$$
$$(\neg\mathsf{Fml}^1_{\mathcal{L}_{\mathsf{T}}^{<n}}(x) \wedge y = x).$$

It remains to verify, in $\mathsf{RA}_n + \mathsf{D}_\mathsf{n}$, that $\mathsf{H}_n(x,y)$ is the required isomorphism from $\mathsf{K_m} \circ \mathsf{M_m}$ to $\mathsf{id}_{\mathsf{RA}_n+\mathsf{D_n}}$, that is, that conditions (2)–(9) on page 9 are satisfied by H_n. D_n, in combination with (34), give us the totality conditions for H. I now verify that $=_i$ and $(\in^i)^{\mathsf{K_n}\circ\mathsf{M_n}}$ behave as expected for $i \leq n$. This will complete the proof.

We first notice that $\mathsf{H}_n(x,y)$ is 'functional', that is, for $m \leq n$:

$$\mathsf{H}_m(x,y) \wedge \mathsf{H}_m(x,z) \to y =_m^{\mathsf{K_m}\circ\mathsf{M_m}} z \tag{36}$$
$$\mathsf{H}_m(x,y) \wedge \mathsf{H}_m(z,y) \to x =_m z. \tag{37}$$

Assuming $\mathsf{H}_i(u,w)$ and $\mathsf{H}_i(x,y)$ with $\mathsf{Fml}_{\mathcal{L}_{\mathsf{T}}^{<i}}(x)$, I first show that

$$\mathsf{RA}_n + \mathsf{D}_n \vdash \mathsf{H}_i(u,w) \wedge \mathsf{H}_i(x,y) \to \big(u =_i^{\mathsf{K_n}\circ\mathsf{M_n}} x \leftrightarrow w =_i y\big).$$

We can safely assume that x, u are $\mathcal{L}_{\mathsf{T}}^{<i}$-formulas. We have

$$\begin{aligned} u =_i^{\mathsf{K_n}\circ\mathsf{M_n}} x &\leftrightarrow u \sim_i^{\mathsf{K_n}} x \\ &\leftrightarrow \forall v(\mathbb{V}_i(\mathsf{sub}^1(v,u)) \leftrightarrow \mathbb{V}_i(\mathsf{sub}^1(v,x))) \\ &\leftrightarrow \forall v(v \in^i w \leftrightarrow v \in^i y) \\ &\leftrightarrow w =_i y \end{aligned}$$

The penultimate line is obtained by definition of H_i. For $\in^i$ I show

$$\mathsf{RA}_n + \mathsf{D}_n \vdash \mathsf{H}_i(u,w) \wedge \mathsf{H}_i(x,y) \to \big(u(\in^i)^{\mathsf{K_n}\circ\mathsf{M_n}}x \leftrightarrow w \in^i y\big)$$

Assuming that $\mathsf{H}_i(u,w)$ and $\mathsf{H}_i(x,y)$ we have, with $\mathsf{Fml}_{\mathcal{L}_{\mathsf{T}}^{<i}}(x)$ and $\mathsf{nu}(u)$:

$$\begin{aligned} u(\in^i)^{\mathsf{K_n}\circ\mathsf{M_n}}x &\leftrightarrow \mathsf{T}_i^{\mathsf{K_n}}(\mathsf{sb}(u,x)) \\ &\leftrightarrow \mathbb{V}_i(\mathsf{sb}(u,x)) \\ &\leftrightarrow u \in^i y \\ &\leftrightarrow w \in^i y. \end{aligned}$$

In the last line we have again employed the properties of H_i. □

COROLLARY 5.11. $\mathsf{RA}_n + \mathsf{D}_n$ *and* RT_n *are e-equivalent.*

Finding principles that render iterated uniform disquotation equivalent to ramifications of parameter free comprehension is, as one might expect, more difficult. At least if one wants resort to principles that represent meaningful restrictions such as the D_n's above. For instance, already to achieve an analogue of Proposition 5.5 by using the same translations, we would require RDT_n to be able to prove a principle like (29). Although RDT_n has full induction, however, there seems to be no way to mimic the crucial role that compositional axioms have in its proof in RT_n. Of course we can add what we lack to RDT_n.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Let

$$\mathsf{E}_n :\leftrightarrow \bigwedge_{i=0}^{n} \forall x\,(\mathsf{T}_i x \leftrightarrow (\exists y : \mathsf{Fml}_{\mathcal{L}_T^{<i}})(\mathbb{T}_i^{\mathsf{M}_n}(y, \mathsf{Ic}^i(x)) \wedge \mathsf{T}_i \mathsf{sb}^1(y, x))).$$

It is somewhat tedious, although straightforward by E_n, to verify that, for $n \in \omega$,

LEMMA 5.12. $\mathsf{RDT}_n + \mathsf{E}_n \vdash \mathsf{E}^{\mathsf{M}_n \circ \mathsf{K}_n}$.

Therefore $\mathsf{RDT}_n + \mathsf{E}_n$ and $\mathsf{RA}_n^{\mathsf{pf}} + \mathsf{E}_n^{\mathsf{K}_n}$ are mutually syntactically embeddable. With Lemma 5.12 at hand, we can conclude:

PROPOSITION 5.13. $\mathsf{RDT}_n + \mathsf{E}_n$ *is a retract of* $\mathsf{RA}_n^{\mathsf{pf}} + \mathsf{E}_n^{\mathsf{K}_n}$.

By adding new principles to the theories in Proposition 5.13, we might even be able to prove the e-equivalence of extensions of $\mathsf{RDT}_n + \mathsf{E}_n$ and $\mathsf{RA}_n^{\mathsf{pf}} + \mathsf{E}_n^{\mathsf{K}_n}$, although the more we add, the more the theories will look convoluted and hardly justifiable.

§6. Conclusion. I have argued that mutual truth-definability, let alone mutual interpretability, is not adequate as sufficient condition for two theories of truth to embody the same conception of truth as defined in §1 and §2. The most prominent example of this failure is represented by case of the well-known principles of truth corresponding to the theories KF and PUTB, that are mutually truth-definable and still do not share most of the distinctive features described in §2. An alternative is the notion of t-equivalence: what I called Thesis 1 claims that it is a sufficient condition for the conceptual equivalence of the notions of truth arising from t-equivalent theories. To support the plausibility of this thesis, already suggested by the relationships between t-equivalence and synonymy given by Friedman and Visser's theorem, I have shown that it captures the manifest non-equivalence of the notions of truth of KF and PUTB—as the theories turn out to be not t-equivalent—and that some intuitively very close notions of truth, such as variations of Tarksian truth and satisfaction in the presence of full induction, result in fact in t-equivalent theories. T-retracts and t-equivalence, therefore, have been devised to be in continuity and extend mutual truth-definability, but if one loosens the criteria on the interpretation of the base-theoretic language, interesting combinations may arise. I have also provided in §4.2 a simple-minded template for generating theories of truth that are bi-interpretable but cannot be t-equivalent.

Mutual truth-definability is a particular case of a mutual syntactical embedding. By following the same strategy adopted to define t-retracts and t-equivalence, one can define the more general notions of e-retract and e-equivalence. These notions have proved to be useful to refine the folklore claims that link iterations of ramified truth and iterations of arithmetical comprehension. By crucially employing recent insights due Albert Visser and Ali Enayat, I have verified that finite iterations of ramified truth are not e-equivalent to finite iterations of arithmetical comprehension; this also shows that the notions of e-retract and e-equivalence are properly stricter than mutual syntactical embeddability. Since typed (disquotational or compositional) truth is an e-retract of finite iterations of arithmetical comprehension, the philosopher interested in deflating the ontological assumptions on the existence of subsets of $\mathbb{N}$ may welcome this phenomenon as a confirmation of this possibility. At the same time, the failure of the t-equivalence between the two hierarchies should suggest that, from the logical point of view, typed truth and membership to (a portion of the) predicatively definable sets are *not* notational variants of each other.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Obviously these are only small, initial steps into the application to theories of truth and subsystems of analysis of the categories of theories and interpretations studied in Visser (2006) and there are a number of unresolved issues. We list a few. In the first place one may consider more examples of natural theories of truth that are mutually truth-definable and investigate their t-equivalence; a natural starting point is, for instance, to consider the hierarchy RT_n for $n \in \omega$ and the corresponding (in the sense of mutual truth-definability) ω-consistent subsets of FS. A second example, related to bi-interpretability simpliciter, concerns theories of truth with the same truth theoretic axioms but built on bi-interpretable base theories: the question is then whether they are uniformly bi-interpretable both in the typed and in the type-free case. In comparing truth axioms and arithmetical comprehension, even more variations are possible: one question is to consider theories with restricted induction (both in the truth-theoretic and the second-order side) that escape Proposition 5.6 and investigate whether they are e-retracts of each other and/or whether they are e-equivalent. Moreover, it is natural to wonder whether it is possible to lift the observations in §5 to transfinite iterations.

§7. Acknowledgments. I thank Ali Enayat, Kentaro Fujimoto, Volker Halbach, Graham Leigh, Lavinia Picollo, Johannes Stern, and the audiences in workshops in Oxford, Paris, Florence, for their feedback. Special thanks go to Albert Visser for allowing me to cite his unpublished arguments, to Thomas Schindler for correcting some inaccuracies in the definitions of the ramified truth theories in §5, and to the anonymous referees for their detailed reports. This work was supported by the Analysis Trust and the European Commission, Grant 658285 FOREMOTIONS.

BIBLIOGRAPHY

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

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

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Enayat, A. & Visser, A. (2015). New constructions of satisfaction classes. In Achourioti, D., Fernández, J. M., Galinon, H., and Fujimoto, K., editors. *Unifying the Philosophy of Truth*. Netherlands: Springer, pp. 321–335.

Enayat, A., Schmerl, J. H., & Visser, A. (2010). ω-models of finite set theory. In Kennedy, J. and Kossak, R., editors. *Set theory, Arithmetic, and Foundations of Mathematics: Theorems, Philosophies*. Lecture Notes in Logic, Vol. 36. La Jolla, CA: Association for Symbolic Logic, 2010, pp. 43–65.

Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, **49**(1), 35–92.

Feferman, S. (1964). Systems of predicative analysis. *The Journal of Symbolic Logic*, **29**(1), 1–30.

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

Feferman, S. (1998). What rests on what? The proof-theoretic analysis of mathematics. In Feferman, S., editor. *In the Light of Logic*. Oxford: Oxford University Press, pp. 187–208.

Field, H. (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Fischer, M. (2009). Minimal truth and interpretability. *The Review of Symbolic Logic*, **2**(4), 799–815.

Fischer, M., Halbach, V., Kriener, J. & Stern, J. (2015). Axiomatizing semantic theories of truth? *The Review of Symbolic Logic*, **8**(2), 257–278.

Friedman, H. & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, **33**, 1–21.

Friedman, H. & A. Visser (2014). *When Bi-Interpretability Implies Synonymy*. Logic Group Preprint Series. University of Utrecht.

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

Hájek P. & Pudlák, P. (1998). *Metamathematics of First-Order Arithmetic*. Berlin: Springer.

Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, **108**, 1–22.

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

Halbach, L. (2014). *Axiomatic Theories of Truth* (revised edition). Cambridge: Cambridge University Press.

Halbach, V. & Horsten, L. (2015). Norms for theories of reflexive truth. In Achourioti, D., Fernández, J. M., Galinon, H. and Fujimoto, K., editors. *Unifying the Philosophy of Truth*. Netherlands: Springer, pp. 263–280.

Heck, R. (2015). Consistency and the theory of truth. *The Review of Symbolic Logic*, **8**(3), 424–466.

Horsten, L. (2011). *The Tarskian Turn: Deflationism and Axiomatic Truth*. Cambridge, MA: MIT Press.

Kaye, R. & Wong, T. L. (2007). On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, **48**(4), 497–510.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, **72**, 690–712.

Laurence, S. & Margolis, E. (1999). *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Leigh, G. E. (2015). Conservativity for theories of compositional truth via cut elimination. *The Journal of Symbolic Logic*, **80**(3), 825–865.

Leigh, G. & Nicolai, C. (2013). Axiomatic truth, syntax, metatheoretic reasoning. *The Review of Symbolic Logic*, **6**(4), 613–636.

Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Blackwell Philosophy Compass* 2. Blackwell, pp. 276–290.

Lutz, S. (2016). What was the syntax-semantics debate in the philosophy of science about? *Philosophy and Phenomenological Research*, forthcoming, doi: 10.1111/phpr.12221.

Nicolai, C. (2016a). A note on typed truth and consistency assertions. *Journal of Philosophical Logic*, **45**(1), 89–119.

Nicolai, C. (2016b). More on systems of truth and predicative comprehension. In Boccuni F. and Sereni A., editors. *Philosophy of Mathematics: Objectivity, Cognition and Proof*. Boston Studies in the History and Philosophy of Science, Netherlands: Springer.

Rogers, H. (1987). *Theory of Recursive Functions and Effective Computability*. Cambridge, MA: MIT Press.

Schwichtenberg, H. & Wainer, S. (2011). *Proofs and Computations*. ASL Lecture Notes Series. Cambridge: Cambridge University Press.

Simpson, S. G. (2009). *Subsystems of Second-Order Arithmetic*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press

Smorynski, C. (1977). The incompleteness theorems. In Barwise, J., editor. *Handbook of Mathematical Logic*. Amsterdam: North Holland, pp. 821–865.
Takeuti, G. (1987). *Proof Theory* (second edition). Amsterdam: North-Holland.
Visser, A. (1991). The formalization of interpretability. *Studia Logica*, **50**(1), 81–105.
Visser, A. (1992). An inside view of Exp. *The Journal of Symbolic Logic*, **57**(1), 131–165.
Visser, A. (1997). An overview of interpretability logic. In Kracht, M., de Rijke, M., and Wansing, H. editors. *Advances in Modal Logic '96*. Stanford, CA: CSLI Publications, pp. 307–359.
Visser, A. (2006). Categories of theories and interpretations. In Enayat, A., Kalantari, I., and Moniri, M., editors. *Logic in Tehran*, Vol. 26. Lecture Notes in Logic, La Jolla, CA: Association for Symbolic Logic, pp. 284–341.
Visser, A. (2015). The interpretability of inconsistency: Feferman's theorem and related results. Forthcoming in the *Bulletin of Symbolic Logic*.
Woodfield, A. (1991). Conceptions *Mind*, **100**(4), 547–572.

MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY
GESCHWISTER-SCHOLL PLATZ 1, MUNICH
GERMANY
E-mail: Carlo.Nicolai@lrz.uni-muenchen.de

https://doi.org/10.1017/S1755020316000435 Published online by Cambridge University Press