J Philos Logic
DOI 10.1007/s10992-017-9434-1

Some Notes on Truths and Comprehension

Thomas Schindler[1]

Received: 27 April 2016 / Accepted: 15 February 2017

Abstract In this paper we study several translations that map models and formulae of the language of second-order arithmetic to models and formulae of the language of truth. These translations are useful because they allow us to exploit results from the extensive literature on arithmetic to study the notion of truth. Our purpose is to present these connections in a systematic way, generalize some well-known results in this area, and to provide a number of new results. Sections 3 and 4 contain some recursion- and proof-theoretic results about Kripke-style fixed-point theories of truth. Section 5 shows how to derive full second-order arithmetic from principles of truth. Section 6 investigates the proof-theoretic strength of disquotation without an arithmetical base theory.

Keywords Truth · Second-order arithmetic · Relative interpretations · Definability

1 Introduction

In this paper we study several translations, essentially known to experts, that map models and formulae of the language of second-order arithmetic to models and formulae of the language of Peano arithmetic augmented by a truth predicate. These

The author wishes to acknowledge and express his appreciation for the support of the Alexander von Humboldt Foundation and the German Research Foundation (Deutsche Forschungsgemeinschaft, "Reference patterns of paradox").

✉ Thomas Schindler
thomas.schindler1980@gmail.com

[1] Clare College, University of Cambridge, Cambridge, UK

Published online: 08 March 2017

translations are useful because they allow us to exploit results from the extensive literature on arithmetic in order to study the notion of truth. Our purpose is to present these connections in a systematic way, to generalize some well-known results in this area, and to provide a number of new results.

While this paper is essentially a technical one, many of its results should be of great philosophical interest. Deflationists such as Horwich [18] claim that truth and satisfaction (truth of) are merely expressive devices. This idea can be spelled out in various ways. According to Parsons [26], the notion of truth enables us to generalize sentence places while the notion of satisfaction allows us to quantify into predicate position. The direct method of generalizing sentence and predicate places is by introducing quantifiable variables that can occupy sentence and predicate position; that is, by introducing second-order variables. A natural way, then, of understanding the idea that truth and satisfaction enable us to generalize sentence respectively predicate places is to say that a theory of truth must be able to emulate or interpret some form of second-order quantification. The translations presented in this paper may help us to make this idea precise.

Secondly, interpretations of (subsystems of) second-order arithmetic into theories of truth can be seen as ontological reductions. In second-order arithmetic, the second-order quantifiers are usually taken to range over sets of numbers. For some sets X we can find a predicate P such that $n \in X$ if and only if P is true of n. Since we can identify predicates with their Gödel numbers, talk of certain sets of numbers can be reduced to or replaced by talk of numbers and satisfaction. Well-known examples in this area are the reduction of the system of arithmetical comprehension to the typed compositional theory of truth and the reduction of ramified analysis to the Tarskian hierarchy of truth. An early philosophical discussion of both results can be found in Parsons [25], who also sketches a proof of both results. These examples show that we may identify arithmetical respectively hyperarithmetical sets with the predicates (or their Gödel codes) of which they are the extension. (For technical details I refer the reader to Halbach [10, 11]. For further discussion, see Halbach [12].) The results in this paper show that we can reduce even stronger subsystems of second-order arithmetic to theories of truth.

There are some further philosophical issues on which our results may shed some light, such as the question of what constitues a good axiomatization of a semantic truth theory (Fischer et al. [8]), or the question of the 'unsubstantiality' of truth (Horsten [17], Shapiro [31], Ketland [19], Halbach [13]). However, I won't attempt to draw any substantial philosophical conclusions in this paper and leave a proper assessment for another occasion.

This paper is structured as follows. In Section 2 we set up the overall framework of the paper. The ideas and results presented there are more or less known to experts. We fix a translation function from the language of second-order arithmetic to the language of truth. The main idea is to translate a formula of the form $t \in Y$ as 'the result of substituting t for the free variable in the formula y is true' or 'y is true of t'. With every interpretation S for the truth predicate we canonically associate a Henkin structure $\mathcal{M}_S$ for the second-order language. It is then shown that a second-order sentence is true in $\mathcal{M}_S$ if and only if its translation is true when S is assigned as the extension of the truth predicate.

In Section 3, we apply the translation lemma to obtain some complexity results for Kripke-style fixed-point theories of truth [20]. We show that whenever a valuation scheme satisfies some minimal criteria, then all inductive sets are weakly definable in the least fixed point of that valuation scheme, while the sets that are strongly definable are exactly the hyperarithmetical ones. The proof presented here is more general than those currently available in the literature as it applies to the Strong Kleene scheme and the supervaluation schemes as well as to Leitgeb's theory [21]. It also applies to a valuation scheme strictly weaker than the Strong Kleene. We also give a very short proof that all inductive sets are definable in Herzberger's revision theory [16].

In Section 4, we draw some proof-theoretic consequences. It is shown that the least fixed points of the valuation schemes satisfying the criteria set out in the previous section validate (the translations of) the theories ID_1 and $\Delta^1_1 - \mathsf{CA}_0$. The axioms of the latter could therefore be taken into account for axiomatizations of the semantic theories. We also present some sufficient conditions under which all fixed points of a valuation scheme satisfy the theory of positive disquotation.

In Section 5 it is shown how to turn comprehension axioms into truth-theoretic axioms. We present a disquotational theory of truth that relatively interprets full second-order arithmetic, and show it to be consistent. This strengthens a result of Schindler [30] that parameter-free second-order arithmetic is reducible to principles of truth.

Section 6 investigates the proof-theoretic strength of uniform disquotation without its base theory. This is done by recovering weak set theories from the T-biconditionals. It is shown that typed uniform disquotation, over logic alone, interprets Tarski's theory R while the theory of positive uniform disquotation, over logic alone, interprets Robinson arithmetic Q.

In Section 7 we conclude with some final remarks.

Technical Preliminaries The language of Peano arithmetic, $\mathcal{L}_{\mathrm{PA}}$, is a first-order language that contains a denumerably infinite set of individual variables $v_0, v_1, v_2, \ldots$, the connectives $\neg, \vee$ and $\wedge$, the quantifiers $\forall, \exists$ and the identity symbol $=$. We assume that all other connectives are defined in the usual way. The sole non-logical symbols are the individual constant $\overline{0}$, the unary function symbol S for the successor function, the binary function symbols $+$ and $\cdot$ for addition and multiplication, respectively, and function symbols for certain primitive recursive (p.r.) functions that we are going to specify in the course of the paper. If h is such a p.r. function, we write $\dot{h}$ (with a subdot) for the corresponding function symbol. The language $\mathcal{L}_{\mathrm{T}}$ is obtained from $\mathcal{L}_{\mathrm{PA}}$ by augmenting the latter with the unary predicate symbol T.

The theory PA contains the defining axioms for zero, successor, addition, multiplication and the other p.r. function symbols together with all instances of the induction axiom scheme

$$\varphi(\overline{0}) \wedge \forall x(\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \varphi(x)$$

where $\varphi(x)$ is a formula of $\mathcal{L}_{\mathrm{PA}}$. The theory PAT is obtained from PA by extending the induction axiom scheme to the full language $\mathcal{L}_{\mathrm{T}}$.

If n is a number, we write $\overline{n}$ for its numeral, i.e., the term that is obtained by applying the successor symbol S n-many times to the constant $\overline{0}$. We assume some

effective Gödelcoding of the expressions of $\mathcal{L}_{\mathrm{T}}$. If σ is some expression, we write $\#\sigma$ for its code and $\ulcorner\sigma\urcorner$ for the numeral of its code. We occasionally identify expressions with their codes.

Let $s_i^k(m,n) = \#\varphi(\overline{n}/x_j)$, provided that $m = \#\varphi$ codes a formula with exactly k free variables and x_j is its i-th free variable (according to the index ordering). If one of these conditions is not satisfied we let $s_i^k(m,n) = 0$. The functions s_i^k are primitive recursive and will be represented by the symbols $\dot{s}_i^k$. Given $\varphi := \varphi(x,y,z)$ with exactly x, y, z free and $\mathrm{index}(x) < \mathrm{index}(y) < \mathrm{index}(z)$, we write $\ulcorner\varphi(\dot{x},\dot{y},\dot{z})\urcorner$ for $\dot{s}_1^1(\dot{s}_2^2(\dot{s}_3^3(\ulcorner\varphi\urcorner, z), y), x)$, and similarly for formulae with n free variables. We often write $\dot{s}$ instead of $\dot{s}_1^1$. Then the uniform T-schema can be written as

$$\forall x_1 \ldots \forall x_n (T\ulcorner\varphi(\dot{x_1} \ldots \dot{x_n}\urcorner) \leftrightarrow \varphi(x_1, \ldots, x_n))$$

For more details on this notation, I refer the reader to Cantini [6] or Halbach [15].

Standard models of $\mathcal{L}_{\mathrm{T}}$ have the form $(\mathbb{N}, S)$, where $\mathbb{N}$ is the standard model of PA and $S \subseteq \omega$ interprets the truth predicate T.

Let us now turn to the language of second-order arithmetic. The language $\mathcal{L}_2$ of second-order arithmetic is obtained from $\mathcal{L}_{\mathrm{PA}}$ by adding the binary relation symbol $\in$ plus second-order or *set variables* $X_0, X_1, X_2, \ldots$ (Let us call $v_0, v_1, \ldots$ *number* variables.) This gives us new formulae of the form $t \in X$ and $\forall X \varphi$.

$\mathcal{L}_2$ is a two-sorted first-order language with usual (first-order) rules for both set and number quantifiers. Intended Henkin models for $\mathcal{L}_2$ have the form $(\mathbb{N}, \mathcal{M})$, where $\mathcal{M} \subseteq \wp(\omega)$ and the set variables X_i range over the elements of $\mathcal{M}$. We call such models ω-models.

A formula $\varphi \in \mathcal{L}_2$ is called *arithmetical* iff it contains no second-order quantifiers. Note that such a formula might contain free second-order variables. (Sometimes we also refer to the formulae of $\mathcal{L}_{\mathrm{PA}}$ as arithmetical. It should always be clear from the context which sense is intended.)

A formula $\varphi \in \mathcal{L}_2$ is Π_n^1 (Σ_n^1) iff its has the form $Q_1 X_1 \ldots Q_n X_n \psi$, where ψ is arithmetical, $Q_1 \ldots Q_n$ is a string of alternating second-order quantifiers, and Q_1 is universal (existential).

A set $X \subseteq \omega$ is Π_n^1 (Σ_n^1) iff $X = \{n \mid (\mathbb{N}, \wp(\omega)) \models \varphi(\overline{n})\}$, where $\varphi(x)$ is a $\Pi_n^1(\Sigma_n^1)$ formula with exactly x free. (That is, a set is Π_n^1 iff it is definable by a Π_n^1-formula in the standard model of second-order arithmetic.) A set $X \subseteq \omega$ is Δ_n^1 iff X is both Π_n^1 and Σ_n^1. For more information on the analytic hierarchy I refer the reader to Shoenfield [32].

The theory Z_2, called *full second-order arithmetic* or *classical analysis*, is formulated in $\mathcal{L}_2$ and contains besides the axioms of PA with induction extended to the full language $\mathcal{L}_2$ all instances of the comprehension axiom scheme

$$\forall y_1 \ldots y_m \, \forall Y_1 \ldots Y_n \, \exists X \, \forall x \, (x \in X \leftrightarrow \varphi(x, y_1, \ldots, y_m, Y_1, \ldots, Y_n))$$

where $\varphi(x, y_1, \ldots, y_m, Y_1, \ldots, Y_n)$ is a formula of $\mathcal{L}_2$ containing exactly the displayed variables free. The subsystem ACA of arithmetical comprehension is obtained by restricting the comprehension axiom scheme to arithmetical $\mathcal{L}_2$-formulae. Subsystems of analysis are studied in Simpson [33].

Another notion that we will use frequently in this paper is that of a *relative interpretation*. Giving a general definition of that notion is a tricky business (cf. Visser

[36]), but for our purposes the following should suffice. Roughly, a theory T_1 in $\mathcal{L}_{T_1}$ is relatively interpretable in a theory T_2 in $\mathcal{L}_{T_2}$ if and only if there is a definitional extension of T_2 in an expansion of $\mathcal{L}_{T_2}$ and a translation I from $\mathcal{L}_{T_1}$ to the expansion of $\mathcal{L}_{T_2}$ such that (i) I replaces every non-logical expression of $\mathcal{L}_{T_1}$ by one of the same kind and arity (but not necessarily the same sort), (ii) I preserves logical structure, possibly relativizing quantifiers, and (iii) I preserves theoremhood. Here, the notion of a definitional extension is the standard one as in e.g. Monk [22, p. 208]; for instance, we assume the usual uniqueness and existence conditions for function symbols. In practice, we will often specify a translation by specifying a formula with n free variables as translation of an n-ary predicate symbol instead of specifying a predicate symbol in an expansion as required by the definition.

We use bold face letters to indicate strings of expressions. For example, $\mathbf{x}$ denotes $x_1, \ldots, x_n$ and $\ulcorner \boldsymbol{\varphi} \urcorner$ denotes $\ulcorner \varphi_1 \urcorner, \ldots, \ulcorner \varphi_n \urcorner$. It should always be clear from the context what the length of the string is.

2 Truth-Sets and Second-Order Structures

In this section we introduce our first translation from $\mathcal{L}_2$ to $\mathcal{L}_{\mathrm{T}}$. The ideas and results presented here are more or less known to experts, although we are not aware of a systematic presentation of them in the literature. The purpose of this section is mainly to set up the overall framework of this paper. Recall that standard models of $\mathcal{L}_{\mathrm{T}}$ have the form $(\mathbb{N}, S)$, where $\mathbb{N}$ interprets the arithmetical vocabulary and $S \subseteq \omega$ interprets the truth predicate T. Let us call S a *truth-set*. Any truth-set $S \subseteq \omega$ encodes or determines a Henkin structure $(\mathbb{N}, \mathcal{M}_S)$ for $\mathcal{L}_2$ as follows:

Definition 1 Let $S \subseteq \omega$ and $\varphi \in \mathit{Form}^1_T$ (=an $\mathcal{L}_{\mathrm{T}}$-formula with exactly one free variable).

1. $S_\varphi = \{n \mid \#\varphi(\overline{n}) \in S\} \subseteq \omega$
2. $\mathcal{M}_S = \left\{S_\varphi \mid \varphi \in \mathit{Form}^1_T\right\} \subseteq \wp(\omega)$

For example, assume that S contains (the codes of) all true $\mathcal{L}_{\mathrm{PA}}$-sentences, i.e., $S \supseteq \{\#\varphi \mid \varphi \in \mathcal{L}_{PA}, \mathbb{N} \models \varphi\}$. Then $\mathcal{M}_S$ contains all arithmetically definable subsets of ω. For instance,

$$S_{x=\overline{3}} = \{n \mid \#(\overline{n} = \overline{3}) \in S\} = \{n \mid \mathbb{N} \models \overline{n} = \overline{3}\} = \{3\}$$

and

$$S_{\exists y(x=\overline{2}y)} = \{n \mid \#(\exists y(\overline{n} = \overline{2}y)) \in S\} = \{0, 2, 4, \ldots\}$$

In particular, for any $S \subseteq \omega$ we have

$$S_{Tx} = \{n \mid \#(T\overline{n}) \in S\} = \{n \mid n \in S\} = S$$

In the terminology of Cantini [5, 6], $\mathcal{M}_S$ is the *envelope* of S; we will call it the set of sets *encoded* by S. Note that by Tarski's undefinability theorem, in general we do *not* have $S_\varphi = \{n \mid (\mathbb{N}, S) \models \varphi(\overline{n})\}$. Hence $\mathcal{M}_S$ does not coincide with the set of sets that are definable in the classical structure $(\mathbb{N}, S)$. However, we will later see

(Section 3) that $\mathcal{M}_S$ may coincide with the set of sets that are definable in S in some non-classical logic.

As mentioned in the introduction, we occasionally identify expressions with their codes. Accordingly, we also write S_k for S_φ, provided that $k = \#\varphi$. Occassionally, we also denote S_φ by S_t when t is a term denoting $\#\varphi$.

Now consider the following translation function from the language $\mathcal{L}_2$ to the truth language $\mathcal{L}_\mathrm{T}$.

Definition 2 The function $^* : \mathcal{L}_2 \to \mathcal{L}_T$ is defined as follows:

$$
\begin{aligned}
&v_i^* = v_{2i} \\
&X_i^* = v_{2i+1} \\
&\overline{0}^* = \overline{0} \\
&(f(t_1, \ldots, t_n))^* = f(t_1^*, \ldots, t_n^*) \\
&(t_1 = t_2)^* = (t_1^* = t_2^*) \\
&(\neg\varphi)^* = \neg\varphi^* \\
&(\varphi \wedge \psi)^* = \varphi^* \wedge \psi^* \\
&(\varphi \vee \psi)^* = \varphi^* \vee \psi^* \\
&(t \in X_i)^* = T\dot{s}(v_{2i+1}, t^*) \\
&(\forall v_i \varphi)^* = \forall v_{2i}\varphi^* \\
&(\exists v_i \varphi)^* = \exists v_{2i}\varphi^* \\
&(\forall X_i \varphi)^* = \forall v_{2i+1}(Fm_T^1(v_{2i+1}) \to \varphi^*) \\
&(\exists X_i \varphi)^* = \exists v_{2i+1}(Fm_T^1(v_{2i+1}) \wedge \varphi^*)
\end{aligned}
$$

Here, the predicate $Fm_T^1(v)$ naturally represents the set of (codes of) formulae of $\mathcal{L}_\mathrm{T}$ that contain exactly one free variable; the function symbol $\dot{s}$ (introduced in the introduction) represents the substitution function defined by $s(\#\varphi, n) = \#\varphi(\overline{n})$, provided φ is a formula with exactly one free variable. Thus, on the above translation, the formula $t \in X_i$ is translated as:

The result of substituting t for the free variable in (the formula) v_{2i+1} is true

The translation replaces set quantifiers with quantifiers relativized to formulae with one free variable; this accords with what we said in the introduction, namely that we want to identify sets of numbers with the predicates of which they are the extension. The set variables are mapped to variables with an odd index, and number variables to variables with an even index. In order to increase readability, we will often omit the indices of the variables and use some suggestive notation instead. For example, the translation of a second-order formula of the form $\forall x \exists Y(x \in Y)$ will simply be written as $\forall x \exists y(Fm_T^1(y) \wedge T\dot{s}(y, x))$.

Sometimes it is convenient to add set constants to $\mathcal{L}_2$. Given $\mathcal{M}_S$, we let the set constant $\overline{S_\varphi}$ denote the set S_φ. We expand our above translation function by letting

$$(t \in \overline{S_\varphi})^* = T\dot{s}(\ulcorner\varphi\urcorner, t^*)$$

If h is a variable assignment for $(\mathbb{N}, \mathcal{M}_S)$, define the assignment h^* for $(\mathbb{N}, S)$ by $h^*(v_{2i}) := h(v_i)$ and $h^*(v_{2i+1}) := \min\{k \in Form_T^1 \mid S_k = h(X_i)\}$. It is easily seen that our translation preserves the denotation of arithmetical terms.

Proposition 1 *Let h be an assignment for $(\mathbb{N}, \mathcal{M}_S)$. Then $t^{(\mathbb{N},\mathcal{M}_S),h} = t^{*(\mathbb{N},S),h^*}$ for all number terms t of $\mathcal{L}_{PA}$.*

The following central proposition shows that a second-order sentence is true in the set of sets encoded by S if and only if its translation into the language of truth is true in the truth-set S (i.e., if the truth predicate is interpreted by S). Roughly, the gist of this result is that we can convert or translate questions regarding truth-sets into questions regarding second-order structures which are often better understood. We will see several applications of this in subsequent sections.

Proposition 2 (Translation Lemma) *Let $S \subseteq \omega$, let $\varphi(\mathbf{y}, \mathbf{X}, \overline{\mathbf{S}_\gamma}) \in \mathcal{L}_2$, and let h be an assignment for $(\mathbb{N}, \mathcal{M}_S)$. Then:*

$$(\mathbb{N}, \mathcal{M}_S), h \models \varphi \Leftrightarrow (\mathbb{N}, S), h^* \models \varphi^*$$

Proof By induction on the complexity of formulae.

The case $t_1 = t_2$ follows from Proposition 1.

Consider $t \in X_i$, where t is any term. Let $t^{(\mathbb{N},\mathcal{M}_S),h} = n$ and $h(X_i) = A$. Then there is a k such that $k = \min\{m \mid S_m = A\}$. Then

$$\begin{aligned}
(\mathbb{N}, \mathcal{M}_S), h \models t \in X_i &\Leftrightarrow n \in A \\
&\Leftrightarrow n \in S_k \\
&\Leftrightarrow s(k, n) \in S \\
&\Leftrightarrow (\mathbb{N}, S) \models Ts(\overline{k}, \overline{n}) \\
&\Leftrightarrow (\mathbb{N}, S), h^* \models Ts(v_{2i+1}, t^*)
\end{aligned}$$

Next, consider $t \in \overline{S_\gamma}$, where t is any term. Let $t^{(\mathbb{N},\mathcal{M}_S),h} = n$. Then

$$\begin{aligned}
(\mathbb{N}, \mathcal{M}_S), h \models t \in \overline{S_\gamma} &\Leftrightarrow n \in S_\gamma \\
&\Leftrightarrow s(\#\gamma, n) \in S \\
&\Leftrightarrow (\mathbb{N}, S) \models Ts(\ulcorner \gamma \urcorner, \overline{n}) \\
&\Leftrightarrow (\mathbb{N}, S), h^* \models Ts(\ulcorner \gamma \urcorner, t^*)
\end{aligned}$$

The cases $\neg\psi$, $\psi \wedge \chi$, $\psi \vee \chi$, $\exists x \psi$ and $\forall x \psi$ follow easily from the I.H.

Now consider $\forall X_i \psi$ and let

$$M = \left\{k \in Form_T^1 \mid \forall m \in Form_T^1 (S_m = S_k \rightarrow k \leq m)\right\}$$

Then:

$$\begin{aligned}
(\mathbb{N}, \mathcal{M}_S), h \models \forall X_i \psi &\Leftrightarrow \forall A \in \mathcal{M}_S : (\mathbb{N}, \mathcal{M}_S), h(A : X_i) \models \psi && (1)\\
&\Leftrightarrow \forall k \in Form_T^1 : (\mathbb{N}, \mathcal{M}_S), h(S_k : X_i) \models \psi && (2)\\
&\Leftrightarrow \forall k \in M : (\mathbb{N}, \mathcal{M}_S), h(S_k : X_i) \models \psi && (3)\\
&\Leftrightarrow \forall k \in M : (\mathbb{N}, S), h^*(k : v_{2i+1}) \models \psi^* && (4)\\
&\Leftrightarrow \forall k \in Form_T^1 : (\mathbb{N}, S), h^*(k : v_{2i+1}) \models \psi^* && (5)\\
&\Leftrightarrow (\mathbb{N}, S), h^* \models \forall v_{2i+1} \left(Fm_T^1(v_{2i+1}) \rightarrow \psi^*\right) && (6)
\end{aligned}$$

The implication from Eqs. 3 to 2 follows from the definition of M and the extensionality of sets. The equivalence between Eqs. 3 and 4 is given by the inductive hypothesis, since $[h(S_k : X_i)]^* = h^*(k : v_{2i+1})$ for every minimal k. The step from Eqs. 4 to 5 is justified because in translated formulae, such as ψ^*, v_{2i+1} occurs only in contexts of the form $T\dot{s}(v_{2i+1}, t)$. By the definition of the sets S_k, if $S_m = S_k$ then

$$(\mathbb{N}, S), h' \models T\dot{s}(\overline{k}, t) \leftrightarrow T\dot{s}(\overline{m}, t),$$

for any term t and assignment h'.

The case $\exists X\psi$ is proved in a similar way. □

As a corollary we get:

Proposition 3 *Let $S \subseteq \omega$ and let φ be a closed $\mathcal{L}_2$-formula. Then:*

$$(\mathbb{N}, \mathcal{M}_S) \models \varphi \Leftrightarrow (\mathbb{N}, S) \models \varphi^*$$

We note the following special case. Let $\varphi(X)$ be an arithmetical $\mathcal{L}_2$-formula that contains X as sole set variable. Then, in any $\mathcal{L}_2$-structure $(\mathbb{N}, \mathcal{M})$, the truth value of φ depends (apart from the numbers assigned to the number variables) only on the set assigned to X. Let P be that set. Then we have

$$(\mathbb{N}, \mathcal{M}), h(P : X) \models \varphi(X) \Leftrightarrow \mathbb{N}, h \models \varphi(\overline{P})$$

where $\overline{P}$ is a set constant interpreted by P. The Translation Lemma implies:

Proposition 4 *Let $\varphi(X)$ be an arithmetical $\mathcal{L}_2$-formula that contains X as sole set variable. Then*

$$\mathbb{N} \models \varphi(\overline{S_k}) \Leftrightarrow (\mathbb{N}, S) \models \varphi^*(t)$$

where $t^{\mathbb{N}} = k$.

Sometimes it is convenient to assume that the language $\mathcal{L}_2$ additionally contains n-ary relation variables $X_1^n, X_2^n, \ldots$ for every $n > 1$. It is straightforward to extend our apparatus to cover this more general situation and we will occasionally make use of this in what follows.

For example, on the syntactic side, a formula of the form $Y^2(x_1, x_2)$ translates into a formula of the form $T\dot{s}(\dot{s}_2^2(y, x_2), x_1)$. A formula of the form $\forall Y^n\varphi$ translates into a formula of the form $\forall y(Fm_T^n(y) \rightarrow \varphi^*)$, where $Fm_T^n(x)$ represents the set of $\mathcal{L}_T$-formulae with exactly n free variables. On the semantic side, if $\varphi(x, y)$ is an $\mathcal{L}_T$-formula containing exactly x, y free, and $S \subseteq \omega$, we let $S_\varphi = \{\langle n, m\rangle \mid \#\varphi(\overline{n}, \overline{m}) \in S\}$. It should be obvious that Propositions 1 to 4 can be extended to the more general case.

Of course, since there is a primitive recursive coding machinery in the language of arithmetic, we could simply do with the unary case. However, I believe that the presentation of some of the later results will be more perspicuous if we avoid additional coding.

3 Definability in Fixed-Point Theories

The Translation Lemma is useful because it allows us to transfer results about arithmetic to theories of truth. Let us illustrate this by proving some definability results for Kripke-style fixed-point theories of truth. These results are well-known for particular choices of a valuation scheme, but the Translation Lemma allows us to generalize these results to a wide range of valuation schemes (in particular, it also applies to Leitgeb's [21] theory of truth and to a scheme strictly weaker than the Strong Kleene, as well as to the revision theory of truth) and prove them in a uniform way. We assume that the reader is familiar with Kripke's theory (cf. Kripke [20]) and introduce the following definitions only to fix some terminology.

A *partial model* for $\mathcal{L}_{\mathrm{T}}$ is a triple $(\mathbb{N}, E, A)$, where $E \subseteq \omega$ is the extension and $A \subseteq \omega$ is the anti-extension of the truth predicate, and $E \cap A = \emptyset$. A valuation scheme V is a function mapping partial models and sentences to the set $\{0, 1, \mathsf{u}\}$. We write $(\mathbb{N}, E, A) \models_V \varphi$ iff V assigns 1 to φ and $(\mathbb{N}, E, A)$. We make a few minimal assumptions:

(V1) The arithmetical vocabulary is interpreted in the standard way and every $\mathcal{L}_{\mathrm{PA}}$-sentence receives the same truth value under V as it receives in the standard model $\mathbb{N}$.
(V2) A sentence of the form Tt receives the value 1 iff $t^{\mathbb{N}} \in E$ and receives the value 0 iff $t^{\mathbb{N}} \in A$.
(V3) A conjunction is true under V iff both conjuncts are true under V.
(V4) A universal sentence is true under V iff all its instances are true under V.
(V5) A sentence receives value 1 under V iff its negation receives value 0 (and vice versa).

In this paper, we consider only valuations V that are monotonic (i.e. V preserves the values 0 and 1 if one moves from a partial model to one extending it). Given a partial model $(\mathbb{N}, E, A)$ and a monotonic valuation scheme V, the Kripke-jump $\mathcal{J}_V$ delivers the partial model $(\mathbb{N}, E', A')$, where E' is the set of sentences that receive value 1 in $(\mathbb{N}, E, A)$ and A' is the set of all sentences that receive value 0 in $(\mathbb{N}, E, A)$. We denote E' by $\mathcal{J}_V(E)$. A partial model $(\mathbb{N}, E, A)$ is a fixed point of $\mathcal{J}_V$ iff $(\mathbb{N}, E, A) = (\mathbb{N}, E', A')$. The least fixed point of $\mathcal{J}_V$ is obtained by iterated applications of the Kripke-jump, starting with the partial model $(\mathbb{N}, \emptyset, \emptyset)$, and taking unions at limit stages. We denote the partial model obtained at stage α by $(\mathbb{N}, E^\alpha, A^\alpha)$, and the least fixed point by $(\mathbb{N}, E^\infty, A^\infty)$. (Note: E^α always refers to the extension of the truth predicate at stage α in the construction of the least fixed point, where we start the hierarchy with the empty set.) The classical *close-off* of an arbitrary partial model $(\mathbb{N}, E, A)$ is the classical model $(\mathbb{N}, E)$.

A set $X \subseteq \omega$ is *weakly definable* in $(\mathbb{N}, E, A)$ iff there is a $\varphi(x) \in \mathcal{L}_T$ such that $X = \{n \mid (\mathbb{N}, E, A) \models_V \varphi(\overline{n})\}$. If $(\mathbb{N}, E, A)$ is a Kripke fixed point, then

$$\begin{aligned} E_\varphi &= \{n \mid \#\varphi(\overline{n}) \in E\} \\ &= \{n \mid (\mathbb{N}, E, A) \models_V T\ulcorner\varphi(\overline{n})\urcorner\} \\ &= \{n \mid (\mathbb{N}, E, A) \models_V \varphi(\overline{n})\} \end{aligned}$$

Thus, whenever $(\mathbb{N}, E, A)$ is a fixed point, $\mathcal{M}_E$ coincides with the collection of sets that are weakly definable in $(\mathbb{N}, E, A)$.

A set $X \subseteq \omega$ is called Π^1_1-*hard* iff every Π^1_1-set Y is many-one reducible to X, i.e., there is a recursive function f such that for all $n \in \omega$, $n \in Y$ iff $f(n) \in X$. A set X is called Π^1_1-*complete* iff X is Π^1_1-hard and X is a Π^1_1-set.

It is well-known that E^∞ is a Π^1_1-complete set of integers for several choices of V. This was proved for the Strong Kleene scheme by Burgess [3] and Cantini [5] (the result was already announced by Kripke), for the van Fraassen scheme by Burgess [3], for the Cantini scheme by Cantini [6],[1] and for Leitgeb's [21] theory (which is based on his notion of semantic dependence) by Welch [40]. The proof in [40] was generalized by Fischer et al. [8] to cover a range of further supervaluational schemes (for example, supervaluations based on maximal consistent extensions). Leitgeb's theory is not defined in the Kripkean way, but in Beringer and Schindler [1] it is shown that Leitgeb's theory coincides with the least fixed point of a particular 3-valued valuation scheme that we dubbed *the Leitgeb valuation scheme*, V_L. Thus, Leitgeb's theory can be considered as a special case of Kripke's.[2] (Thus, whenever we talk about Kripke's theory of truth in what follows, Leitgeb's theory is always included.) We observe the following:

Proposition 5 *If $\mathcal{M}_{E^\infty}$ contains all Π^1_1-sets, then E^∞ is Π^1_1-hard.*

Proof For every Π^1_1-set P we need to find a recursive function such that $n \in P$ iff $f(n) \in E^\infty$. By assumption, $P = (E^\infty)_\varphi \in \mathcal{M}_{E^\infty}$ for some $\varphi(x)$. Let $f(n) := \#\varphi(\overline{n})$. □

Using the Translation Lemma, we can give a short and uniform proof that E^∞ is Π^1_1-hard for a wide range of valuation schemes, including all of the schemes mentioned above. (We call the class of these valuation schemes *nice*.) There is no general argument by which one could show that E^∞ is itself a Π^1_1-set, but for the valuation schemes that are usually discussed in the truth-theoretic literature such arguments are straightforward.

The general strategy of the proof is as follows. Any Π^1_1-set P can be seen as the slice or projection of the fixed point I of a particular operator. This operator is given by a particular formula φ of $\mathcal{L}_2$. Whenever V is nice, we can utilize our translation function to find a formula φ' of $\mathcal{L}_\mathrm{T}$ such that $(E^\infty)_{\varphi'} = I$. The set P can be easily

[1] By the van Fraassen scheme we mean the scheme vF introduced in Burgess [3] and by the Cantini scheme we mean the scheme FV introduced in Cantini [6].

[2] We briefly sketch Leitgeb's theory for those readers not familiar with it. A sentence φ depends on $S \subseteq \omega$ iff for all $P \subseteq \omega$ we have $(\mathbb{N}, P) \models \varphi \Leftrightarrow (\mathbb{N}, S \cap P) \models \varphi$. Let $G_0 = \emptyset$ and let $G_{\alpha+1}$ be the set of (codes of) sentences that depend on G_α. If γ is a limit ordinal, let G_γ be the union of all G_α for $\alpha < \gamma$. Next, let $T_0 = \emptyset$ and let $T_{\alpha+1}$ be the set of (codes of) sentences that are elements of G_α and are true in the classical model $(\mathbb{N}, T_\alpha)$. At limit stages, we take unions again. This hierarchy reaches a fixed point, T^∞, which Leitgeb proposes as interpretation for the truth predicate. The valuation scheme V_L is defined as follows. Let $V_L(\mathbb{N}, E, A)(\varphi) = 1$ if φ depends on $E \cup A$ and $(\mathbb{N}, E) \models \varphi$, and $= 0$ if φ depends on $E \cup A$ and $(\mathbb{N}, E) \not\models \varphi$, and $= \mathsf{u}$ otherwise. Then $E^\infty = T^\infty$.

recovered from $(E^{\infty})_{\varphi'}$. This shows that $\mathcal{M}_{E^{\infty}}$ contains all Π^1_1-sets and we can apply Proposition 5.

We briefly recall some concepts and results from the theory of inductive definitions. For more details we refer the reader to Moschovakis [23]. Suppose that $\varphi(x_1, \ldots, x_n, X^n)$ is an arithmetical $\mathcal{L}_2$-formula (with all free variables displayed) in which X^n occurs only *positively* (that is, every occurrence of X^n is in the scope of an even number of negation signs). The positive elementary operator $\Gamma_\varphi : \wp(\omega^n) \to \wp(\omega^n)$ determined by φ is defined by

$$\Gamma_\varphi(S) = \{\langle k_1, \ldots, k_n\rangle \mid \mathbb{N} \models \varphi(\overline{k}_1, \ldots, \overline{k}_n, \overline{S})\}$$

This operator is monotone in the sense that, whenever $S \subseteq S'$, then $\Gamma(S) \subseteq \Gamma(S')$. Let us inductively define

- $I^0_\varphi = \emptyset$
- $I^{\alpha+1}_\varphi = \Gamma_\varphi(I^\alpha_\varphi)$
- $I^\gamma_\varphi = \bigcup_{\alpha<\gamma} I^\alpha_\varphi$, when γ is a limit ordinal.

Let $I_\varphi := I^\kappa_\varphi$ where κ is least with $I^\kappa_\varphi = I^{\kappa+1}_\varphi$. We call I_φ the least fixed point of the positive elementary operator Γ_φ, or simply the set build up by Γ_φ. The existence of I_φ follows from the monotonicity of the operator Γ_φ and a simple cardinality argument.

It is a well-known result that if P is a Π^1_1-set then there is a formula $\varphi := \varphi(u, y, X^2)$ such that for all n, $n \in P$ iff $\langle\langle\rangle, n\rangle \in I_\varphi$. In other words, P is the $\langle\rangle$-slice of I_φ. Here, $\varphi(u, y, X^2)$ is of the form

$$Seq(u) \wedge (\psi(u, y) \vee \forall z X^2(u^\frown z, y)) \tag{7}$$

for some $\mathcal{L}_{\mathrm{PA}}$-formula $\psi(u, y)$, $Seq(u)$ expresses that u is the code of a (finite) sequence of natural numbers, $u^\frown z$ denotes the concatenation of the sequence u with the sequence $\langle z\rangle$, and $\langle\rangle$ denotes the empty sequence.

Our goal is to find a formula φ' of $\mathcal{L}_{\mathrm{T}}$ such that $(E^{\infty})_{\varphi'} = I_\varphi$. In fact, we will show that $(E^\alpha)_{\varphi'} = I^\alpha_\varphi$ for every α.

To this end, let I^*_φ be a closed term of $\mathcal{L}_{\mathrm{T}}$ such that

$$I^*_\varphi = \ulcorner \varphi^*(u, y, I^*_\varphi)\urcorner$$

is provable in PA. Such a term can be constructed by diagonalisation.[3] The definition of the term I^*_φ is adapted from Cantini [5]. Unpacking notation, we observe that $\varphi^*(u, y, I^*_\varphi)$ is shorthand for

$$Seq(u) \wedge (\psi(u, y) \vee \forall z T\dot{s}(\dot{s}^2_2(I^*_\varphi, y), u^\frown z)$$

The displayed formula is our desired formula φ'.

[3] Let $\iota\varphi := \ulcorner \varphi^*(v_1, v_2, \dot{s}^3_3(v_3, v_3))\urcorner$ and $I^*_\varphi := \dot{s}^3_3(\iota\varphi, \iota\varphi)$ Observe that

$$\begin{aligned} I^*_\varphi &= \dot{s}^3_3(\iota\varphi, \iota\varphi) \\ &= \ulcorner \varphi^*(v_1, v_2, \dot{s}^3_3(\iota\varphi, \iota\varphi))\urcorner \\ &= \ulcorner \varphi^*(v_1, v_2, I^*_\varphi)\urcorner \end{aligned}$$

There are some valuation schemes such that E^∞ is *not* Π^1_1-hard (for instance, Cain and Damnjanovic [4] have shown that the Weak Kleene scheme does not necessarily generate a Π^1_1-hard fixed point, depending on the chosen Gödel coding).[4] However, we will show that whenever V is nice in the sense of the following definition, E^∞ will indeed be Π^1_1-hard.

Definition 3 A valuation scheme V is *nice* iff the following conditions hold for all partial models $(\mathbb{N}, E, A)$ and sentences $\varphi, \psi \in \mathcal{L}_T$:

(N1) if $\psi \in \mathcal{L}_{PA}$ and $\mathbb{N} \models \psi$ then $(\mathbb{N}, E, A) \models_V \psi \vee \varphi$
(N2) if $(\mathbb{N}, E, A) \models_V \varphi$ and $\psi \in \mathcal{L}_{PA}$ then $(\mathbb{N}, E, A) \models_V \psi \vee \varphi$
(N3) if a disjunction $\psi \vee \varphi$ is true under V and $\psi \in \mathcal{L}_{PA}$, then (at least) one of ψ, φ is true under V

In other words, the conditions require that if $\psi \in \mathcal{L}_{PA}$, then $\psi \vee \varphi$ is true under V iff ψ or φ is true under V. The Strong Kleene scheme, the Leitgeb scheme, the van Fraassen scheme and the Cantini scheme are all nice. Moreover, it is easily seen that any supervaluation scheme is nice.[5] The Weak Kleene scheme, however, is not nice, because it does not satisfy property (N1). There is a nice valuation scheme strictly lying between the Weak and Strong Kleene scheme. It is obtained by adopting the Weak Kleene rules for the existential and universal quantifier but the Strong Kleene rules for conjunction and disjunction.

Properties (N1)-(N3) ensure (together with (V1)-(V4)) that the formula $\varphi^*(u, y, I^*_\varphi)$ is satisfied in a partial model $(\mathbb{N}, E, A)$ if and only if it is satisfied in its classical close-off $(\mathbb{N}, E)$:

Proposition 6 *Assume that V is nice and that $\varphi(u, y, X^2)$ is an instance of Eq. 7. Then for all $m, n \in \omega$ and all partial models $(\mathbb{N}, E, A)$ we have:*

$$(\mathbb{N}, E, A) \models_V \varphi^*(\overline{m}, \overline{n}, I^*_\varphi) \Leftrightarrow (\mathbb{N}, E) \models \varphi^*(\overline{m}, \overline{n}, I^*_\varphi)$$

Proof $\Rightarrow$: Assume $(\mathbb{N}, E, A) \models_V \varphi^*(\overline{m}, \overline{n}, I^*_\varphi)$, that is

$$(\mathbb{N}, E, A) \models_V Seq(\overline{m}) \wedge (\psi(\overline{m}, \overline{n}) \vee \forall z T \dot{s}(\dot{s}^2_2(I^*_\varphi, y), u^\frown z))$$

Since the sentence in question is a conjunction, (V3) implies that both conjuncts must hold in $(\mathbb{N}, E, A)$. Since the first conjunct is arithmetical, (V1) implies that it must be true in the standard model, hence it also holds in the classical model $(\mathbb{N}, E)$. The second conjunct is a disjunction, of which the first disjunct is again arithmetical. Thus by (N3), either $\psi(\overline{m}, \overline{n})$ or $\forall z T \dot{s}(\dot{s}^2_2(I^*_\varphi, y), u^\frown z)$ must be true in the partial model

[4] For additional results on the Weak Kleene scheme see Speranski [34], where it is shown that adding a symbol for proper substraction solves the problem.

[5] Where a scheme is supervaluational iff it is given by a rule of the following form: $(\mathbb{N}, E, A) \models_V \varphi$ iff for all $S \supseteq E$, if $\Phi(S, E, A)$ then $(\mathbb{N}, S) \models \varphi$, where Φ is a condition such that for every partial model $(\mathbb{N}, E, A)$ there is an $S \supseteq E$ with $\Phi(S, E, A)$.

$(\mathbb{N}, E, A)$. In either case, the disjunction will hold in the classical model $(\mathbb{N}, E)$ as well.

$\Leftarrow$: Assume $(\mathbb{N}, E) \models \varphi^*(\overline{m}, \overline{n}, I_\varphi^*)$, that is

$$(\mathbb{N}, E) \models Seq(\overline{m}) \wedge (\psi(\overline{m}, \overline{n}) \vee \forall z T \dot{s}(\dot{s}_2^2(I_\varphi^*, y), u^\frown z))$$

By (V3) it suffices to show that both conjuncts hold in $(\mathbb{N}, E, A)$. We already know that it holds in its close-off. Since the first conjunct is arithmetical, it holds in the partial model because of (V1). So let us turn to the second conjunct (the disjunction). Now, if $(\mathbb{N}, E) \models \psi(\overline{m}, \overline{n})$, then, since ψ is arithmetical, it must be true in the standard model of arithmetic and therefore, by (V1), hold in the partial model as well. But then the disjunction must hold in the partial model because of property (N1) of a nice valuation. On the other hand, suppose that $(\mathbb{N}, E) \models \forall z T \dot{s}(\dot{s}_2^2(I_\varphi^*, y), u^\frown z)$. Then this must hold in the partial model because of (V1), (V2) and (V4). But then (N2) ensures that the disjunction must hold there as well. □

Combining the above proposition with the Translation Lemma we get:

Proposition 7 *Assume that V is nice and that $\varphi(u, y, X^2)$ is an instance of Eq. 7. Then for all $\alpha \in ON$ we have:*

$$I_\varphi^\alpha = (E^\alpha)_{I_\varphi^*} := \left\{\langle m, n\rangle \mid \#\varphi^*\left(\overline{m}, \overline{n}, I_\varphi^*\right) \in E^\alpha\right\}$$

In other words, the set I_φ^α build up by the positive operator Γ_φ at stage α is identical with the set weakly defined by the formula $\varphi^*(u, y, I_\varphi^*)$ in the partial model obtained in the Kripke construction at stage α (starting with the empty set as extension and anti-extension at stage 0).

Proof By transfinite induction on α.

$\alpha = 0$: Since $E^0 = \emptyset$, we get $(E^0)_{I_\varphi^*} = \emptyset = I_\varphi^0$.

$\alpha = \beta + 1$: Let $\langle m, n\rangle \in (E^{\beta+1})_{I_\varphi^*}$, whence by definition $\#\varphi^*(\overline{m}, \overline{n}, I_\varphi^*) \in E^{\beta+1}$. Since (at successor levels) the extension of the truth predicate comprises precisely those sentences that are true at the preceding level, we get $(\mathbb{N}, E^\beta, A^\beta) \models_V \varphi^*(\overline{m}, \overline{n}, I_\varphi^*)$. It follows from Proposition 6 that the formula also holds in the classical close-off, that is $(\mathbb{N}, E^\beta) \models \varphi^*(\overline{m}, \overline{n}, I_\varphi^*)$. By I.H., $(E^\beta)_{I_\varphi^*} = I_\varphi^\beta$, so Proposition 4 yields $\mathbb{N} \models \varphi\left(\overline{m}, \overline{n}, \overline{I_\varphi^\beta}\right)$, whence by definition $\langle m, n\rangle \in I_\varphi^{\beta+1}$.

For the other direction, assume that $\langle m, n\rangle \in I_\varphi^{\beta+1}$. So $\mathbb{N} \models \varphi\left(\overline{m}, \overline{n}, \overline{I_\varphi^\beta}\right)$, whence by I.H. and Proposition 4, $(\mathbb{N}, E^\beta) \models \varphi^*(\overline{m}, \overline{n}, I_\varphi^*)$. Hence, by Proposition 6, $(\mathbb{N}, E^\beta, A^\beta) \models_V \varphi^*(\overline{m}, \overline{n}, I_\varphi^*)$ and therefore $\#\varphi^*(\overline{m}, \overline{n}, I_\varphi^*) \in E^{\beta+1}$ by definition of Kripke jump, which means that $\langle m, n\rangle \in (E^{\beta+1})_{I_\varphi^*}$.

α is a limit ordinal: by I.H. and definition we get:

$$(E^\alpha)_{I_\varphi^*} = \bigcup_{\beta<\alpha} (E^\beta)_{I_\varphi^*} = \bigcup_{\beta<\alpha} I_\varphi^\beta = I_\varphi^\alpha$$

□

Now we can easily see that all Π^1_1-sets are weakly definable in the minimal fixed points of a nice valuation scheme.

Proposition 8 *If V is nice then*

1. $\mathcal{M}_{E^\infty} \supseteq \{P \mid P \text{ is } \Pi^1_1\}$
2. E^∞ *is* Π^1_1*-hard*

Proof Ad 1. If P is a Π^1_1-set, then for appropriate φ we have $n \in P$ iff $\langle\langle\rangle, n\rangle \in I_\varphi$ for every n. Let $\varphi' := \varphi^*(\langle\rangle, x, I^*_\varphi)$. It follows from the previous theorem that $P = (E^\infty)_{\varphi'}$.

Ad 2. This follows from the first item and Proposition 5. □

Note also the following:

Proposition 9 *If V is nice and E^∞ is itself Π^1_1, then E^∞ is Π^1_1-complete and $\mathcal{M}_{E^\infty} = \{P \mid P \text{ is } \Pi^1_1\}$.*

Proof It suffices to show that $\mathcal{M}_{E^\infty} \subseteq \{P \mid P \text{ is } \Pi^1_1\}$. Now if $(E^\infty)_\varphi \in \mathcal{M}_{E^\infty}$, then $(E^\infty)_\varphi$ is elementary definable in the structure $(\mathbb{N}, E^\infty)$ by the formula $Ts(\ulcorner\varphi\urcorner, x)$ and thus is Π^1_1, because (by assumption) E^∞ is Π^1_1. □

Note that in order to prove the above complexity results it is sufficient to assume that (N1)-(N3) hold merely for those partial models that arise in the construction of the least fixed point.

Let us call an $\mathcal{L}_{\mathrm{T}}$-formula $\varphi(x)$ *total with respect to* S (or S-total) iff

$$(\mathbb{N}, S) \models \forall x(T\ulcorner\varphi(\dot{x})\urcorner \vee T\ulcorner\neg\varphi(\dot{x})\urcorner)$$

That is, a formula is S-total iff for every n, either $\#\varphi(\overline{n}) \in S$ or $\#\neg\varphi(\overline{n}) \in S$. If $\varphi(x)$ is E^∞-total, negation clause (V5) implies that every instance $\varphi(\overline{n})$ will have a definite truth value in the partial model $(\mathbb{N}, E^\infty, A^\infty)$.

Let us say that a set $X \subseteq \omega$ is *strongly definable* in the partial model $(\mathbb{N}, E, A)$ iff $X = \{n \mid (\mathbb{N}, E, A) \models_V \varphi(\overline{n})\}$ for some E-total formula φ. It is known for the Strong Kleene and Cantini scheme that the sets strongly definable in the least Kripke fixed point are exactly the hyperarithmetical sets. (See Cantini [5, 6].) The proposition below extends this result to all valuation schemes that are nice.

A set P is *hyperarithmetical* iff both P and its complement are Π^1_1 (i.e. if P is Δ^1_1). It is well-known that a set P is hyperarithmetical iff $P = J_i$ for some H-index i, where J_i and H are inductively defined as follows (see e.g. Shoenfield [32, chapter 7.9]). Let W_e denote the domain of the e-th partial recursive function under some enumeration W.

- For each e, $(0, e)$ is an H-index.
- If e is an H-index, then $(1, e)$ is an H-index.
- If every number in W_e is an H-index, then $(2, e)$ is an H-index.

For each H-index i we define a set J_i as follows.

- If $i = (0, e)$, then $J_i = W_e$.
- If $i = (1, e)$ and $e \in H$, then $J_i = (\omega \setminus J_e)$.
- If $i = (2, e)$ and $W_e \subseteq H$, then $J_i = \bigcup_{k \in W_e} J_k$.

Thus, the hyperarithmetical sets are obtained by starting with the recursively enumerable sets and closing them under complements and certain countable unions. Let *HYP* denote the collection of hyperarithmetical sets.

Proposition 10 *If V is nice and E^∞ is Π^1_1, then*

$$HYP = \mathcal{M}^{tot}_{E^\infty} := \{(E^\infty)_\varphi \mid \varphi \text{ is } E^\infty\text{-total}\}$$

Proof "$\subseteq$": With every set J_i we will associate an E^∞-total $\mathcal{L}_{\mathrm{T}}$-formula φ_i such that $J_i = (E^\infty)_{\varphi_i}$. Notice first that for every W_e there is an $\mathcal{L}_{\mathrm{PA}}$-formula $\psi_e(x)$ that elementary defines W_e in $\mathbb{N}$. Now, if $i = (0, e)$ let $\varphi_i = \psi_e(x)$. If $i = (1, e)$ then let $\varphi_i = \neg\varphi_e(x)$. If $i = (2, e)$ then let $\varphi_i = \forall y(\neg\psi_e(y) \vee T\dot{s}(y, x))$. Using the Translation Lemma and the properties of a nice valuation scheme, one proves by induction on the H-indices that every φ_i is E^∞-total and $(E^\infty)_{\varphi_i} = J_i$. If $i = (0, e)$ this follows from (V1). If $i = (1, e)$ this follows from the I.H. and the definition of E^∞-totality. If $i = (2, e)$ this follows from the I.H. and (N1) and (V4).

"$\supseteq$": Assume that φ is E^∞-total. We have to show that $(E^\infty)_\varphi$ is Δ^1_1. Under the assumptions, Proposition 9 implies that both $(E^\infty)_\varphi$ and $(E^\infty)_{\neg\varphi}$ are Π^1_1. Now by (V5), $\neg\varphi$ is E^∞-total too, and $(E^\infty)_{\neg\varphi}$ is the complement of $(E^\infty)_\varphi$. This means that $(E^\infty)_\varphi$ must be Σ^1_1. Thus $(E^\infty)_\varphi$ is both Σ^1_1 and Π^1_1, hence it is Δ^1_1. □

Notice that the only use of (V5) that we made in this paper is in the $\supseteq$-direction of the proof of Proposition 10. All previous results (and most results in the next section) can be proved without assuming (V5).

At the beginning of this Section 1 briefly mentioned that it is also possible to show that every Π^1_1-set is weakly definable in the revision theory of truth. (Of course, it is well-known that the revision theory defines much more sets than that. See Welch [39].) In fact, the proof is even simpler than for Kripke's theory. Let us briefly consider the case of Herzberger's theory in [16].

Let $R_0 = \emptyset$, let $R_{\alpha+1} = \{\#\varphi \mid (\mathbb{N}, R_\alpha) \models \varphi\}$ and let

$$R_\gamma = \{\#\varphi \mid \exists\beta < \gamma \forall\alpha(\beta \leq \alpha < \gamma \Rightarrow \#\varphi \in R_\alpha\}$$

when γ is a limit ordinal.

Proposition 11 *Let Γ_φ be an elementary positive operator given by the formula $\varphi(x_1, \ldots, x_n, X^n)$. Let $I^*_\varphi = \ulcorner\varphi^*(\boldsymbol{x}, I^*_\varphi)\urcorner$. Then $(R_\alpha)_{I^*_\varphi} = I^\alpha_\varphi$ for every $\alpha \leq \kappa$ (where κ is the closure ordinal of the operator Γ_φ).*

Notice that Proposition 11 is more general than Proposition 7: here, φ can be any elementary X^n-positive formula, in contrast to Proposition 7 which only applies to formulae of the form Eq. 7. In the next section, we will see that this strengthening also holds for Kripke-style theories that satisfy stronger requirements than (N1)-(N3).

Proof By transfinite induction on α. Trivial for $\alpha = 0$. Let $\alpha = \beta + 1$. Then

$$\begin{aligned}
(R_{\beta+1})_{I^*_\varphi} &= \{\mathbf{m} \mid \#\varphi^*(\overline{\mathbf{m}}, I^*_\varphi) \in R^{\beta+1}\}, && \text{by definition}\\
&= \{\mathbf{m} \mid (\mathbb{N}, R_\beta) \models \varphi^*(\overline{\mathbf{m}}, I^*_\varphi)\}, && \text{by definition}\\
&= \{\mathbf{m} \mid \mathbb{N} \models \varphi(\overline{\mathbf{m}}, \overline{(R_\beta)_{I^*_\varphi}})\}, && \text{by prop. 4}\\
&= \left\{\mathbf{m} \mid \mathbb{N} \models \varphi(\overline{\mathbf{m}}, \overline{I^\beta_\varphi})\right\}, && \text{by I.H.}\\
&= I^{\beta+1}_\varphi, && \text{by definition}
\end{aligned}$$

Finally, for limits γ, if $\mathbf{m} \in (R_\gamma)_{I^*_\varphi}$ then by definition there must be a β such that $\varphi^*(\mathbf{m}, I^*_\varphi)$ is true in all models $(\mathbb{N}, R^\alpha)$ where $\beta \leq \alpha < \gamma$. Thus, by I.H. and Proposition 4, $\mathbb{N} \models \varphi(\overline{\mathbf{m}}, I^\alpha_\varphi)$ for $\beta \leq \alpha < \gamma$, which implies $\mathbf{m} \in I^\gamma_\varphi$.

Conversely, if $\mathbf{m} \in I^\gamma_\varphi$ then there is some β such that for all α with $\beta \leq \alpha < \gamma$, $\mathbf{m} \in I^\alpha_\varphi$. By I.H. and Proposition 4 we get $\mathbf{m} \in (R_\alpha)_{I^*_\varphi}$ for all α with $\beta \leq \alpha < \gamma$ and thus $\mathbf{m} \in (R_\gamma)_{I^*_\varphi}$. Therefore $(R_\gamma)_{I^*_\varphi} = I^\gamma_\varphi$. □

4 Axioms

The correspondence between truth-sets and second-order models that we introduced in Section 2 seems to be a good way to measure the amount of second-order quantification that a semantic theory of truth is able to mimick. The theorems of the previous section indicate, I believe, a certain lower bound on the proof-theoretic strength that we should expect from a good axiomatization of the minimal Kripke fixed points. For example, a semantic theory that encodes all inductive sets should be axiomatized by a theory that formalizes or interprets the theory of inductive definitions, ID_1. A similar idea is suggested in Fischer et al. [8, section 3.2].

The language $\mathcal{L}_{\mathrm{ID}_1}$ extends the language $\mathcal{L}_{\mathrm{PA}}$ by a predicate constant $\overline{I}_\varphi$ for every *arithmetical* $\mathcal{L}_2$-formula $\varphi(x, X)$ (with all free variables displayed) in which X occurs only positively. On the intended interpretation, the constant $\overline{I_\varphi}$ is interpreted by the least fixed point I_φ of the operator Γ_φ.

ID_1 is the theory in $\mathcal{L}_{\mathrm{ID}_1}$ that contains in addition to the axioms of PA and full induction in $\mathcal{L}_{\mathrm{ID}_1}$ all sentences of the form

$$\forall x(\varphi(x, \overline{I_\varphi}) \rightarrow \overline{I_\varphi}(x))$$

and

$$\forall x(\varphi(x, \psi) \rightarrow \psi(x)) \rightarrow \forall x(\overline{I_\varphi}(x) \rightarrow \psi(x))$$

where $\psi(x)$ is an arbitrary formula of $\mathcal{L}_{\mathrm{ID}_1}$ containing exactly x free, and $\varphi(x, \psi)$ is obtained from $\varphi(x, X)$ by replacing every occurrence of $t \in X$ by $\psi(t)$. In order to avoid variable clashes, we assume that variables in ψ are renamed if necessary. The first axiom expresses that $\overline{I_\varphi}$ is closed under the operator defined by the formula $\varphi(x, X)$ while the second expresses that $\overline{I_\varphi}$ is the least such set. For more on ID_1, see Pohlers [27].

The system $\widehat{\mathsf{ID}}_1$ is the subtheory of ID_1 that contains besides the axioms of PA and full induction in $\mathcal{L}_{\mathrm{ID}_1}$ all sentences of the form

$$\forall x(\overline{I_\varphi}(x) \leftrightarrow \varphi(x, \overline{I_\varphi}))$$

where $\varphi(x, X)$ and $\overline{I_\varphi}$ are as above. This axiom expresses that $\overline{I_\varphi}$ is a fixed point of the operator Γ_φ (but not necessarily the least one).

Since every fixed point I_φ of an elementary positive operator Γ_φ is Π^1_1, we can find (via Proposition 8) a formula $\varphi' \in \mathcal{L}_T$ such that $I_\varphi = (E^\infty)_{\varphi'}$ for every nice valuation V. (Note, however, that while φ is an arbitrary X-positive formula, φ' will be of the form Eq. 7.) In order to translate ID_1 into $\mathcal{L}_\mathrm{T}$, we may stipulate that $\overline{I_\varphi}$ is translated as $\ulcorner\varphi'\urcorner$ and $\overline{I_\varphi}(x)$ as $T\dot{s}(\ulcorner\varphi'\urcorner, x)$. Call the resulting translation $^\#$. (On the arithmetical vocabulary, $^\#$ is just the identity function.) As I will point out below, the translation $^\#$ has a certain disadvantage, but it is enough for the following result, which is a corollary of Proposition 8:

Proposition 12 *If V is nice then* $(\mathbb{N}, E^\infty) \models (\mathsf{ID}_1)^\#$

For example, if V is the Strong Kleene scheme, then V satisfies the antecedent of the above proposition. The Kripke-Feferman theory KF (see Feferman [7]) can be seen as an axiomatization of the Strong Kleene fixed points, but it doesn't count as a satisfactory axiomatization of the minimal fixed point if one adheres to the proof-theoretic criterion suggested at the beginning of this section. Burgess has given a variant of KF, called KFB (the acronym stands for 'Kripke-Feferman-Burgess') that does have the same strength as ID_1. The additional strength over KF is obtained by adding a minimality axiom scheme which basically says that if $\varphi(x)$ satisfies the KF axioms, then all true sentences fall under the extension of $\varphi(x)$ (cf. Halbach [15, chap. 17] for details). Another way to strengthen KF, which has already been pointed out by Cantini [5, p. 105], is to add the (translation of the) second axiom scheme of ID_1 to KF. Proposition 12 above shows that the resulting system is still sound with respect to the minimal Strong Kleene fixed point.

The theory PUTB (the acronym stands for *positive uniform T-biconditionals*) is a subtheory of KF and has been investigated by Cantini [5] and Halbach [14]. In the following paragraphs, we prove some simple facts about models of PUTB. In Section 6 we return to this theory, where we will investigate the proof-theoretic strength of positive disquotation without the arithmetic base theory.

Definition 4 PUTB is the theory in $\mathcal{L}_\mathrm{T}$ that extends PAT by all sentences of the form

$$\forall x_1 \ldots \forall x_n (T\ulcorner\varphi(\dot{x_1}, \ldots, \dot{x_n})\urcorner \leftrightarrow \varphi(x_1, \ldots, x_n))$$

where φ is T-positive (i.e. every occurrence of T is in the scope of an even number of negation signs).

The theory PUTB interprets the theory $\widehat{\mathsf{ID}}_1$, but the translation $^\#$ introduced above won't do. Under the translation $^\#$, a fixed point I_φ gets mapped to a formula φ' with the appropriate extension, but the formula φ' does not 'preserve' the syntactic shape of the formula φ. (To repeat, the formula φ' is an instance of Eq. 7, while φ can be an arbitrary X-positive formula.) In order to get a formula with the right syntactic properties, we apply Cantini's trick again. Thus, where $\varphi(x, X)$ is a

X-positive formula, let I^*_φ be a closed term of $\mathcal{L}_\mathrm{T}$ such that $I^*_\varphi = \ulcorner\varphi^*(x, I^*_\varphi)\urcorner$. Now expand the translation * (of Section 2) by translating $\overline{I_\varphi}$ as I^*_φ.

Proposition 13 (Cantini [5]) $\mathsf{PUTB} \vdash (\widehat{\mathsf{ID}}_1)^*$

Proof Since $\varphi(x, X)$ is X-positive, $\varphi^*(x, I^*_\varphi)$ is T-positive, hence the following is an axiom of PUTB:

$$\forall x(T\ulcorner\varphi^*(\dot{x}, I^*_\varphi)\urcorner \leftrightarrow \varphi^*(x, I^*_\varphi))$$

Since $I^*_\varphi = \ulcorner\varphi^*(x, I^*_\varphi)\urcorner$, this is equivalent to

$$\forall x(T\dot{s}(I^*_\varphi, x) \leftrightarrow \varphi^*(x, I^*_\varphi))$$

The preceding formula is the translation of

$$\forall x(\overline{I_\varphi}(x) \leftrightarrow \varphi(x, \overline{I_\varphi}))$$

(Re-naming variables where appropriate.) □

Notice that the proof doesn't work if we use the translation $^\#$ instead of *, because $\ulcorner\varphi'\urcorner \neq \ulcorner\varphi^*(x, \ulcorner\varphi'\urcorner)\urcorner$.

We have already seen that the close-offs of minimal fixed points of nice valuation schemes satisfy $(\mathsf{ID}_1)^\#$ and therefore $(\widehat{\mathsf{ID}}_1)^\#$. But do all of them also satisfy PUTB and therefore $(\widehat{\mathsf{ID}}_1)^*$? If V is strong in the sense of the definition given below, the answer is 'yes'. In that case, the answer applies not only to the minimal fixed points but to arbitrary fixed points as well. Moreover, if V is strong in the sense of the following definition then the minimal fixed point also satisfies $(\mathsf{ID}_1)^*$.

In order to state the definition, we need to fix some terminology. A formula of $\mathcal{L}_T$ is in negation normal form if it is build up from atomic and negated atomic formulas using $\wedge$, $\vee$, $\forall$ and $\exists$ only, without further use of $\neg$. For every formula $\varphi \in \mathcal{L}_T$, let φ^{nf} be a unique formula in negation normal form that is logically equivalent (in classical logic) to φ. To fix things, we let φ^{nf} be the unique formula that is obtained from φ by applying the transformation rules in [2, chap. 19].

Definition 5 A valuation scheme V is *strong* iff the following conditions hold for all partial models $(\mathbb{N}, E, A)$ and sentences $\varphi, \psi \in \mathcal{L}_T$:

(S1) if φ is true under V, then $\varphi \vee \psi$ and $\psi \vee \varphi$ are true under V
(S2) if for some t, $\varphi(t)$ is true under V, then $\exists x\varphi$ is true under V
(S3) if $(\mathbb{N}, E, A) \models_V \varphi$ then $(\mathbb{N}, E) \models \varphi$
(S4) $(\mathbb{N}, E, A) \models_V \varphi$ iff $(\mathbb{N}, E, A) \models_V \varphi^{\mathrm{nf}}$

Note that (S1) implies (N1) and (N2). Condition (S3) expresses that V is *classically sound*: if a sentence is true in a partial model, then it is true in its classical close-off. The Strong Kleene, the van Fraassen and the Cantini scheme are strong. The Weak Kleene and the Leitgeb scheme are not strong, however, because both

violate condition (S1).[6] In Beringer and Schindler [1] it is shown that the Cantini scheme is the strongest valuation scheme that is classically sound. Thus, no supervaluational scheme stronger than Cantini's satisfies condition (S3). Properties (S1)-(S4) enable us to extend Proposition 6 to every T-positive formula:

Proposition 14 *Let* $(\mathbb{N}, E, A)$ *be a partial model and let* φ *be a* T*-positive formula. If* V *is strong then*

$$(\mathbb{N}, E, A) \models_V \varphi \Leftrightarrow (\mathbb{N}, E) \models \varphi$$

Proof The left-to-right direction follows from the classical soundness of V, (S3). The right-to-left direction is proved by induction on the build-up of the T-positive formula φ. We assume that φ is in negation normal form. This is justified by (S4). If φ is of the form $s = t$, $s \neq t$, or Tt, this is trivial. Since $\neg Tt$ is not T-positive, we can ignore it. If φ is a disjunction, the claim follows from the induction hypothesis and property (S1) of a strong valuation. Similarly, if φ is a conjunction, a universal or existential statement, the claim follows from the induction hypotheses and properties (V3), (V4) and (S2), respectively. □

Proposition 15 *If* V *is strong and* $(\mathbb{N}, E, A)$ *is any fixed point of* $\mathcal{J}_V$ *then* $(\mathbb{N}, E) \models$ PUTB.

Proof Let φ be a T-positive sentence and assume that $(\mathbb{N}, E) \models \varphi$. Then Proposition 14 shows that $(\mathbb{N}, E, A) \models_V \varphi$, which implies by the fixed point property that $(\mathbb{N}, E, A) \models_V T^{\ulcorner}\varphi^{\urcorner}$. This means $\#\varphi \in E$, so also $(\mathbb{N}, E) \models T^{\ulcorner}\varphi^{\urcorner}$. Now assume that $(\mathbb{N}, E) \models \neg\varphi$. Since V is classically sound (S4), φ has value 0 or u in the partial model $(\mathbb{N}, E, A)$ and consequently, $\#\varphi \notin E$. So $(\mathbb{N}, E) \models \neg T^{\ulcorner}\varphi^{\urcorner}$. So $(\mathbb{N}, E) \models \varphi \leftrightarrow T^{\ulcorner}\varphi^{\urcorner}$ for all T-positive φ, whence by standardness $(\mathbb{N}, E) \models$ PUTB. □

I do not know the answer to the following:

Question 1 If $(\mathbb{N}, E, A)$ is the minimal (any) fixed point of the Leitgeb valuation scheme, do we have $(\mathbb{N}, E) \models$ PUTB?

A further consequence of Proposition 14 is the following:

Proposition 16 *Suppose that* V *is strong. Let* $\varphi(x_1, \ldots, x_n, X^n)$ *be an arithmetical* $\mathcal{L}_2$*-formula (with all free variables displayed) in which* X^n *occurs only positively, and* $I^*\varphi = {}^{\ulcorner}\varphi^*(\boldsymbol{x}, I^*_\varphi)^{\urcorner}$*. Then* $(E^\alpha)_{I^*_\varphi} = I^\alpha_\varphi$ *for all* $\alpha \in ON$.

The difference to Proposition 7 is that here φ can be an arbitrary arithmetical X^n-positive formula while Proposition 7 only applied to formulae of the form Eq. 7.

[6] This claim is obvious for the Weak Kleene scheme. For the Leitgeb scheme, a counter-example to (S1) can be found in Schindler [29].

Proof By transfinite induction on α. Trivial if $\alpha = 0$. Let $\alpha = \beta + 1$. Then we have:

$$\begin{aligned}
(E^{\beta+1})_{I^*_\varphi} &= \{\mathbf{m} \mid \#\varphi^*(\overline{\mathbf{m}}, I^*_\varphi) \in E^{\beta+1}\}, && \text{by definition}\\
&= \{\mathbf{m} \mid (\mathbb{N}, E^\beta, A^\beta) \models_V \varphi^*(\overline{\mathbf{m}}, I^*_\varphi)\}, && \text{by Kripke jump}\\
&= \{\mathbf{m} \mid (\mathbb{N}, E^\beta) \models \varphi^*(\overline{\mathbf{m}}, I^*_\varphi)\}, && \text{by Proposition 14}\\
&= \{\mathbf{m} \mid \mathbb{N} \models \varphi(\overline{\mathbf{m}}, \overline{(E^\beta)_{I^*_\varphi}})\}, && \text{by Proposition 4}\\
&= \left\{\mathbf{m} \mid \mathbb{N} \models \varphi(\overline{\mathbf{m}}, \overline{I^\beta_\varphi})\right\}, && \text{by I.H.}\\
&= I^{\beta+1}_\varphi, && \text{by definition}
\end{aligned}$$

In the third line, Proposition 14 applies because the translation of a positive $\mathcal{L}_2$-formula is T-positive.

If α is a limit ordinal, the claim follows easily from the I.H. □

As a corollary to Proposition 16, we get:

Proposition 17 *If V is strong then* $(\mathbb{N}, E^\infty) \models (\mathsf{ID}_1)^*$

In the previous section, we have seen that the sets that are strongly definable in the least fixed points of nice valuation schemes are exactly the hyperarithmetical sets. A little modification of the Translation Lemma shows that the system $\Delta^1_1 - \mathsf{CA}_0$ is true in the classical close-offs of those fixed points.

Δ^1_1-CA_0 is the theory in $\mathcal{L}_2$ that contains in addition to the axioms of Q and axioms of arithmetical comprehension all sentences of the form

$$\forall \mathbf{Y} \forall \mathbf{y} (\forall x (\varphi(x, \mathbf{y}, \mathbf{Y}) \leftrightarrow \psi(x, \mathbf{y}, \mathbf{Y})) \rightarrow \exists X \forall x (x \in X \leftrightarrow \varphi(x, \mathbf{y}, \mathbf{Y}))),$$

where $\varphi(x, \mathbf{y}, \mathbf{Y}) \in \mathcal{L}_2$ is a Π^1_1-formula and $\psi(x, \mathbf{y}, \mathbf{Y}) \in \mathcal{L}_2$ is a Σ^1_1-formula, and the induction axiom

$$\forall X (0 \in X \wedge \forall x (x \in X \rightarrow x + \overline{1} \in X) \rightarrow \forall x (x \in X)).$$

The minimal ω-model of the system Δ^1_1-CA_0 is the structure $(\mathbb{N}, HYP)$. (For details, I refer the reader to Simpson [33].)

Let $tot(x)$ abbreviate the formula

$$Fm^1_T(x) \wedge \forall y (T\dot{s}(x, y) \vee T\dot{s}(\dot{\neg} x, y))$$

Here, $\dot{\neg}$ represents the function that sends the code of a formula to the code of its negation. The function $^{**} : \mathcal{L}_2 \rightarrow \mathcal{L}_T$ is defined as the function * except for the following clause:

$$(\forall X_i \varphi)^{**} = \forall v_{2i+1} (tot(v_{2i+1}) \rightarrow \varphi^{**})$$

Let $\mathcal{M}^{tot}_S = \{S_\varphi \mid \varphi \in \mathcal{L}_T, \varphi \text{ is } S - \text{total}\}$. The following proposition states that whenever φ is a second-order formula, then its translation under ** is true in S iff the original sentence is true in $\mathcal{M}^{tot}_S$.

Proposition 18 *Let $S \subseteq \omega$ and $\varphi(\mathbf{y}, \boldsymbol{X}, \overline{\boldsymbol{S}_\gamma}) \in \mathcal{L}_2$. Then:*

$$\left(\mathbb{N}, \mathcal{M}^{tot}_S\right) \models \varphi \Leftrightarrow (\mathbb{N}, S) \models \varphi^{**}$$

Proof Similar to the proof of Proposition 2. □

Proposition 19 *If V is nice and E^∞ is Π^1_1 then*

$$(\mathbb{N}, E^\infty) \models \left(\Delta^1_1\text{-}\mathsf{CA}_0\right)^{**}$$

Proof Since $\mathcal{M}^{tot}_{E^\infty} = HYP$ by Proposition 10 and $(\mathbb{N}, HYP) \models \Delta^1_1 - \mathsf{CA}_0$, the claim follows from Proposition 18. □

5 Reducing Comprehension to Disquotation

The results in the last section suggest that we can obtain proof-theoretically strong disquotational theories of truth by adopting as axioms translations of comprehension axioms. More precisely, the idea is that we take the following uniform T-biconditionals as axioms:

$$\forall \mathbf{z} \forall \mathbf{y} \left(Fm^1_T(\mathbf{y}) \rightarrow \forall x (T^\ulcorner \varphi^*(\dot{x}, \dot{\mathbf{z}}, \dot{\mathbf{y}})^\urcorner \leftrightarrow \varphi^*(x, \mathbf{z}, \mathbf{y})) \right) \tag{8}$$

where $\varphi(x, \mathbf{z}, \mathbf{Y})$ is an arbitrary formula of $\mathcal{L}_2$, and $Fm^1_T(\mathbf{y})$ is shorthand for $Fm^1_T(y_1) \wedge \ldots \wedge Fm^1_T(y_n)$.

Unfortunately, this will lead to inconsistency. The reason is that we can instantiate the quantifier on $\forall y$ to *any* $\mathcal{L}_\mathrm{T}$-formula, in particular to formulae that are not in the range of the translation function *, such as the liar sentence.

Proposition 20 *The schema* (8) *is inconsistent with* PA.

Proof Consider the $\mathcal{L}_2$-formula $\neg(x \in Y)$. This is arithmetical, and its translation is $\neg T\dot{s}(y, x)$. (Let's assume that the index of the variable y is higher than that of x.) Applying (8) to $\neg T\dot{s}(y, x)$ and unpacking notation we get

$$\forall y (Fm^1_T(y) \rightarrow \forall x (T\dot{s}(\dot{s}^2_2(^\ulcorner \neg T\dot{s}(y, x)^\urcorner, y), x) \leftrightarrow \neg T\dot{s}(y, x)))$$

Let $\overline{n}$ be $^\ulcorner \neg T\dot{s}(z, z)^\urcorner$. Since PA proves $Fm^1_T(\overline{n})$, we get:

$$\begin{aligned}
\forall x (T\dot{s}(\dot{s}^2_2(^\ulcorner \neg T\dot{s}(y, x)^\urcorner, \overline{n}), x) &\leftrightarrow \neg T\dot{s}(\overline{n}, x)) \\
\forall x (T\dot{s}(^\ulcorner \neg T\dot{s}(\overline{n}, x)^\urcorner, x) &\leftrightarrow \neg T\dot{s}(\overline{n}, x)) \\
T\dot{s}(^\ulcorner \neg T\dot{s}(\overline{n}, x)^\urcorner, \overline{n}) &\leftrightarrow \neg T\dot{s}(\overline{n}, \overline{n}) \\
T^\ulcorner \neg T\dot{s}(\overline{n}, \overline{n})^\urcorner &\leftrightarrow \neg T\dot{s}(\overline{n}, \overline{n}) \\
T^\ulcorner \neg T\dot{s}(\overline{n}, \overline{n})^\urcorner &\leftrightarrow \neg T^\ulcorner \neg T\dot{s}(\overline{n}, \overline{n})^\urcorner
\end{aligned}$$

The last biconditional follows from the previous one because $\dot{s}(\overline{n}, \overline{n}) = {}^\ulcorner \neg T\dot{s}(\overline{n}, \overline{n})^\urcorner$. □

There are at least two ways out of the problem. First, one could restrict the permissible instances of Eq. 8 to formulae that do not contain free set variables. The resulting system is ω-consistent and interprets Z_2^-, that is second-order arithmetic

with comprehension for all $\mathcal{L}_2$-formulae that do not contain free set variables. This was shown in Schindler [30]. The second option is to keep the parameters but restrict the quantifers in Eq. 8 to a suitable class of formulae. The obvious choice is to restrict the quantifiers to formulae that lie in the range of the translation function; thus, we may instantiate the quantifier on y above to the term $\ulcorner \neg T\dot{s}(y, x)\urcorner$, for example, but not to $\ulcorner \neg T\dot{s}(z, z)\urcorner$.

We will first show how this works for the case of arithmetical $\mathcal{L}_2$-formulae because the general case involves some subtleties that obscure the main idea for the consistency proof. At the end of this section, we will sketch how to extend the method to formulae of arbitrary complexity. It should be possible to extend it to third- and higher-order systems as well. Since we will only deal with classical theories in what follows, we assume for simplicity that the symbols $\vee, \exists$ are no longer among the primitive vocabulary but defined.

We first expand the language of $\mathcal{L}_2$ by adding abstraction terms. This is necessary to close the range of the translation function under instantiations of quantified formulae (see below). Let $\mathcal{L}_2^0$ be the result of adding to $\mathcal{L}_2$ a term $\{x \mid \varphi\}$ for every $\mathcal{L}_{\mathrm{PA}}$-formula φ containing only the number variable x free. Let $\mathcal{L}_2^{n+1}$ be the result of adding to $\mathcal{L}_2^n$ a term $\{x \mid \varphi\}$ for every *arithmetical* $\mathcal{L}_2$-formula φ containing only the number variable x free. Thus, while φ must not contain any set variables at all, it may contain expressions of the form $t \in \{x \mid \psi\}$ as subformulae (where ψ is an arithmetical formula of $\mathcal{L}_2^n$). Finally, let $\mathcal{L}_2^+$ be the union of these languages.

Proposition 21 *The recursion theorem for primitive recursive functions yields the existence of a primitive recursive translation function $\tau : \mathcal{L}_2^+ \to \mathcal{L}_T$ such that:*

$$\tau(\sigma) = \begin{cases} x_{2i}, & \text{if } \sigma \text{ is } x_i \\ x_{2i+1}, & \text{if } \sigma \text{ is } X_i \\ f(\tau(t_1), \ldots, \tau(t_n)), & \text{if } \sigma \text{ is } f(t_1, \ldots, t_n) \\ \ulcorner \tau(\varphi)\urcorner, & \text{if } \sigma \text{ is } \{x_i \mid \varphi\} \\ \tau(t_1) = \tau(t_2), & \text{if } \sigma := t_1 = t_2 \\ T\dot{s}(\tau(t_2), \tau(t_1)) & \text{if } \sigma := t_1 \in t_2 \\ \neg\tau(\psi) & \text{if } \sigma := \neg\psi \\ \tau(\psi) \wedge \tau(\chi) & \text{if } \sigma := \psi \wedge \chi \\ \forall x_{2i} \tau(\psi) & \text{if } \sigma := \forall x_i \psi \\ \forall x_{2i+1}(\exists y(x_{2i+1} = \dot{\tau}(y) \wedge Fm_T^1(x_{2i+1})) \to \tau(\psi)) & \text{if } \sigma := \forall X_i \psi \end{cases}$$

where $\dot{\tau}$ is a function symbol for τ in $\mathcal{L}_T$.

We abbreviate the formula $\exists y(x = \dot{\tau}(y) \wedge Fm_T^1(x))$—which expresses that x is in the range of the translation function τ—by $Trsl(x)$.

The reason for adding the abstraction terms to the language $\mathcal{L}_2$ is to ensure that the predicate $Trsl(x)$ is closed under "instantiation" in the following sense: if $\varphi(x_{2i+1})$ is in the range of τ and $t = \ulcorner \tau(\psi)\urcorner$ for some ψ in the domain of τ, then $\varphi(t/x_{2i+1})$ is also in the range of τ. Note that this is provable in PA (by "internal" induction on the complexity of φ). This is crucial for the proof of Proposition 22 below.

Definition 6 The theory UTB(ACA) is given by the axioms of PAT plus all instances of the following scheme:

$$\forall \mathbf{z} \forall \mathbf{y}(Trsl(\mathbf{y}) \rightarrow \forall x(T^{\ulcorner}\tau(\varphi)(\dot{x}, \dot{\mathbf{z}}, \dot{\mathbf{y}})^{\urcorner} \leftrightarrow \tau(\varphi)(x, \mathbf{z}, \mathbf{y}))) \tag{9}$$

where $\varphi(x, \mathbf{z}, \mathbf{Y}) \in \mathcal{L}_2$ is arithmetical, and $Trsl(\mathbf{y})$ is shorthand for $Trsl(y_1) \wedge \ldots \wedge Trsl(y_n)$.

It is well known that ACA, which is non-conservative over PA, is interpretable in the compositional theory CT but not in the disquotational theory TB, which is a conservative extension of PA (for a proof, see Halbach [15]). The following proposition shows that ACA is interpretable in the disquotational theory UTB(ACA). While the latter (presumably) does not prove any (non-trivial) truth-theoretic generalizations for its own truth predicate, it is able (via the interpretability result) to define a truth predicate for PA for which the compositional clauses are derivable.

Proposition 22 ACA *is relatively interpretable in* UTB(ACA).

Proof It is easy to see that the translations of all arithmetical and induction axioms of ACA are derivable in PAT. Let $\varphi \in \mathcal{L}_2$. We have to show that the translation of the comprehension axiom for φ is a theorem of UTB(ACA). We instantiate the quantifiers in Eq. 9 and unpack notation to get:

$$Trsl(\mathbf{y}) \rightarrow \forall x(T\dot{s}(^{\ulcorner}\tau(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})^{\urcorner}, x) \leftrightarrow \tau(\varphi)(x, \mathbf{z}, \mathbf{y}))$$

Since PA proves that $^{\ulcorner}\tau(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})^{\urcorner}$ is the code of a formula of $\mathcal{L}_{\mathrm{T}}$ that has exactly x free, and τ is "closed under instantiations" (provably in PA), we have

$$Trsl(\mathbf{y}) \rightarrow Trsl(^{\ulcorner}\tau(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})^{\urcorner})$$

This implies:

$$Trsl(\mathbf{y}) \rightarrow (Trsl(^{\ulcorner}\tau(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})^{\urcorner}) \wedge \forall x(T\dot{s}(^{\ulcorner}\tau(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})^{\urcorner}, x) \leftrightarrow \tau(\varphi)(x, \mathbf{z}, \mathbf{y})))$$

and, by logic:

$$Trsl(\mathbf{y}) \rightarrow \exists v(Trsl(v) \wedge \forall x(T\dot{s}(v, x) \leftrightarrow \tau(\varphi)(x, \mathbf{z}, \mathbf{y})))$$

Now we re-introduce universal quantifiers:

$$\forall \mathbf{z} \forall \mathbf{y}((Trsl(\mathbf{y}) \rightarrow \exists v(Trsl(v) \wedge \forall x(T\dot{s}(v, x) \leftrightarrow \tau(\varphi)(x, \mathbf{z}, \mathbf{y}))))$$

This is the translation of the comprehension axiom for φ. □

In order to show the consistency of UTB(ACA), we will define a set $S \subseteq \omega$ such that $\mathcal{M}_S$ is the collection *ARITH* of arithmetically definable sets. Since $(\mathbb{N}, ARITH)$ is a model of ACA, a variant of the Translation Lemma will then show that $(\mathbb{N}, S)$ is a model of UTB(ACA).

Let the term $\{x \mid \varphi\}$ denote the set $\{n \mid (\mathbb{N}, ARITH) \models \varphi(\overline{n})\} \in ARITH$. This is well-defined, since *ARITH* is closed under arithmetical definability.

Now we let

1. $S = \{\#\tau(\varphi) \mid \varphi \in \mathcal{L}_2^+, (\mathbb{N}, ARITH) \models \varphi\}$

2. $S_{\tau(\varphi)} = \{n \mid \#\tau(\varphi(\overline{n})) \in S\}$
3. $\mathcal{M}_S = \left\{S_{\tau(\varphi)} \mid \varphi \in \mathcal{L}_2^+\right\}$

It is not very hard to see that $\mathcal{M}_S = ARITH$. By induction on the build-up of φ we prove:

Proposition 23 *For every* $\varphi \in \mathcal{L}_2^+$*:* $(\mathbb{N}, ARITH) \models \varphi \Leftrightarrow (\mathbb{N}, S) \models \tau(\varphi)$

This is proved in a similar way as Proposition 2. The only new case is $t \in \{x \mid \varphi\}$. Recall that $\{x \mid \varphi\}$ denotes the set of n such that $(\mathbb{N}, ARITH) \models \varphi(\overline{n})$. Let $t^{\mathbb{N}} = m$. Then

$$\begin{aligned}(\mathbb{N}, ARITH) \models t \in \{x \mid \varphi\} &\Leftrightarrow (\mathbb{N}, ARITH) \models \varphi(\overline{m}) \\ &\Leftrightarrow \#\tau(\varphi(\overline{m})) \in S \\ &\Leftrightarrow (\mathbb{N}, S) \models T\dot{s}(\ulcorner\tau(\varphi)\urcorner, \tau(t))\end{aligned}$$

Proposition 24 *The theory* UTB(ACA) *has an* ω*-model.*

Proof Let $\varphi(x, \mathbf{z}, Y_1, \ldots, Y_k)$ be an arithmetical $\mathcal{L}_2$-formula with exactly the displayed variables free. Let $n, \mathbf{m} \in \omega$ be arbitrary and $r_1, \ldots, r_k \in \omega$ be such that $\mathbb{N} \models Trsl(\overline{r}_i)$ for $i \leq k$. Then $r_i = \#\tau(\psi_i)$ for some ψ_i. Thus

$$\begin{aligned}&(\mathbb{N}, S) \models \tau(\varphi)(\overline{n}, \overline{\mathbf{m}}, \ulcorner\tau(\psi_1)\urcorner, \ldots, \ulcorner\tau(\psi_k)\urcorner) \\ \Leftrightarrow\ &(\mathbb{N}, ARITH) \models \varphi(\overline{n}, \overline{\mathbf{m}}, \{x \mid \psi_1\}, \ldots, \{x \mid \psi_k\}) \quad \text{by Prop. 23} \\ \Leftrightarrow\ &\#\tau(\varphi)(\overline{n}, \overline{\mathbf{m}}, \ulcorner\tau(\psi_1)\urcorner, \ldots, \ulcorner\tau(\psi_k)\urcorner) \in S \quad \text{by defn of } S \\ \Leftrightarrow\ &(\mathbb{N}, S) \models T\ulcorner\tau(\varphi)(\overline{n}, \overline{\mathbf{m}}, \ulcorner\tau(\psi_1)\urcorner, \ldots, \ulcorner\tau(\psi_k)\urcorner)\urcorner\end{aligned}$$

Since $(\mathbb{N}, S)$ is a standard model, we have

$$\forall \mathbf{z} \forall \mathbf{y}(Trsl(\mathbf{y}) \rightarrow \forall x(T\dot{s}(\ulcorner\tau(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})\urcorner, x) \leftrightarrow \tau(\varphi)(x, \mathbf{z}, \mathbf{y})))$$

Thus, $(\mathbb{N}, S) \models$ UTB(ACA). □

Now let us see how this method can be extended to $\mathcal{L}_2$-formulae of arbitrary complexity. In order to prove the consistency of the resulting system, we will have to define a set $S \subseteq \omega$ such $(\mathbb{N}, \mathcal{M}_S)$ is a model of second-order arithmetic. Pick a *countable* set $\mathcal{A} \subseteq \wp(\omega)$ such that $(\mathbb{N}, \mathcal{A})$ is a model of Z_2. Such models exist; we refer the interested reader to Putnam et al. [28]. That $(\mathbb{N}, \mathcal{A})$ is a model of Z_2 implies that $\mathcal{A}$ is closed under second-order definabilty with parameters from $\mathcal{A}$.

Since our goal is to construct S such that $\mathcal{M}_S = \mathcal{A}$, every set in $\mathcal{A}$ must be coded by a formula (with one free variable) of $\mathcal{L}_{\mathrm{T}}$. In contrast to the arithmetical case, it is not clear that every member of $\mathcal{A}$ can be defined by a formula of $\mathcal{L}_2$. Thus we need to expand $\mathcal{L}_2$ by countably many set constants $P_1, P_2, \ldots$; then we inductively add an abstraction term $\{x \mid \varphi\}$ for *every* formula of the expanded language that contains exactly one free number variable. Call the resulting language $\mathcal{L}_2^\dagger$.

Now we define a translation τ' from $\mathcal{L}_2^\dagger$ to $\mathcal{L}_{\mathrm{T}}$ as in Proposition 21. In order to translate the set constants, we pick for every i a term P_i^* with $P_i^* = \ulcorner T\dot{s}(P_i^*, v_{2i})\urcorner$

(via diagonalization). The translation function τ' is defined exactly like τ with the proviso that $\tau'(P_i) = P_i^*$ and $\tau'(t \in P_i) = T\dot{s}(P_i^*, \tau'(t))$.

Definition 7 The theory $\mathsf{UTB}(\mathsf{Z}_2)$ is given by the axioms of PAT plus all instances of the following scheme:

$$\forall \mathbf{z} \forall \mathbf{y}(Trsl'(\mathbf{y}) \rightarrow \forall x(T\dot{s}(\ulcorner \tau'(\varphi)(x, \dot{\mathbf{z}}, \dot{\mathbf{y}})\urcorner, x) \leftrightarrow \tau'(\varphi)(x, \mathbf{z}, \mathbf{y}))) \quad (10)$$

where $\varphi(x, \mathbf{z}, \mathbf{Y}) \in \mathcal{L}_2$.

Similarly to Proposition 22, we can prove:

Proposition 25 Z_2 *is relatively interpretable in* $\mathsf{UTB}(\mathsf{Z}_2)$.

The inclusion of the terms P_i^* into the range of the translation function is entirely dispensable for the interpretability result, but they play a crucial role in the consistency proof.

Let us now indicate how to modify the proof of Proposition 24 to obtain a consistency proof for $\mathsf{UTB}(\mathsf{Z}_2)$. Let e be a bijection between $\{P_1, P_2, \ldots\}$ and $\mathcal{A}$. Such a function exists since $\mathcal{A}$ is countable. We let the constant P_i denote the set $e(P_i)$. Moreover, we let the term $\{x \mid \varphi\}$ denote the set $\{n \mid (\mathbb{N}, \mathcal{A}) \models \varphi(\overline{n})\}$. This is well-defined, since $\mathcal{A}$ is closed under definability.

As above, we let S consist of the set of sentences that are translations of $\mathcal{L}_2^\dagger$-sentences that hold in $(\mathbb{N}, \mathcal{A})$.

Obviously, $\mathcal{M}_S \subseteq \mathcal{A}$. Moreover, since every set in $\mathcal{A}$ is definable by a formula of the form $x \in P_i$, we can find an $\mathcal{L}_{\mathrm{T}}$-formula φ (namely, the translation of $x \in P_i$ under τ') such that $S_\varphi = e(P_i)$. This implies $\mathcal{A} \subseteq \mathcal{M}_S$, and therefore $\mathcal{A} = \mathcal{M}_S$. A variant of the Translation Lemma yields the desired consistency proof.

Proposition 26 *The theory* $\mathsf{UTB}(\mathsf{Z}_2)$ *has an* ω*-model.*

While $\mathsf{UTB}(\mathsf{Z}_2)$ is by no means a very elegant theory, it shows that principles of truth need not be any weaker than our strongest arithmetical theories. The expressive power gained by second-order quantification can be recovered in theories of truth. There is, however, a certain mismatch. While second-order languages allows us to generalize over any predicate of the second-order language, the truth predicate of $\mathsf{UTB}(\mathsf{Z}_2)$ does not allow us to generalize over all predicates of the language of truth, but only over those predicates that can be translated back into the second-order formalism.

6 Truth Without a Base Theory

In the last section, we saw how to derive comprehension axioms from uniform T-biconditionals. The derivation made use of some of the axioms of Peano arithmetic. The use of those arithmetical axioms was essential, because we needed to show that the formula in question is in the range of the translation function. This, in turn, is

crucial because, under our translation, we relativized the set quantifiers to formulae of a particular shape. If, on the other hand, we do not relativize the set quantifiers to a predicate when setting up the translation, no arithmetical axioms are needed for deriving comprehension axioms from uniform T-biconditionals.

In order to see this more clearly, let $\mathcal{L}_{\in}$ be the first-order language with identity whose sole non-logical constant is the binary relation symbol '$\in$'. The language $\mathcal{L}_{\in}$ may be regarded as a sublanguage of $\mathcal{L}_{\mathrm{T}}$ by stipulating that $x \in y$ is defined as, or is shorthand for $T\dot{s}_1^1(y, x)$. Now, it is easy to see that any given instance of the comprehension axiom scheme,

$$\forall x_1 \ldots \forall x_n \, \exists y \, \forall x_0 \, (x_0 \in y \leftrightarrow \varphi(x_0, x_1, \ldots, x_n)),$$

where y is not free in φ, is derivable *in pure first-order logic* from the corresponding instance of the uniform T-schema,

$$\forall x_1 \ldots \forall x_n \forall x_0 \, (T\ulcorner \varphi(\dot{x_0}, \dot{x_1}, \ldots, \dot{x_n})\urcorner \leftrightarrow \varphi(x_0, x_1, \ldots, x_n))$$

While the uniform T-schema is still formulated in the language of $\mathcal{L}_{\mathrm{T}}$ (its formulation involves the substitution function symbols and the Gödel numerals), the proof can be carried out in pure first-order logic in the sense that no (non-logical) axioms of PA are used in the derivation. For, by convention, $T\ulcorner \varphi(\dot{x_0}, \dot{x_1}, \ldots, \dot{x_n})\urcorner$ is shorthand for

$$T\dot{s}_1^1 \left(\dot{s}_2^2(\ldots \dot{s}_{n+1}^{n+1}(\ulcorner \varphi \urcorner, x_n \right), \ldots, x_1), x_0)$$

Thus, instantiating the quantifiers $\forall x_1 \ldots \forall x_n$ we get

$$\forall x_0 \left(T\dot{s}_1^1\left(\dot{s}_2^2 \left(\ldots \dot{s}_{n+1}^{n+1}(\ulcorner \varphi \urcorner, x_n), \ldots, x_1\right), x_0\right) \leftrightarrow \varphi(x_0, x_1, \ldots, x_n)\right),$$

whence by existential weakening,

$$\exists y \forall x_0 \left(T\dot{s}_1^1(y, x_0) \leftrightarrow \varphi(x_0, x_1, \ldots, x_n)\right)$$

Now we can re-introduce the universal quantifers and get

$$\forall x_1 \ldots \forall x_n \, \exists y \, \forall x_0 \, \left(T\dot{s}_1^1(y, x_0) \leftrightarrow \varphi(x_0, x_1, \ldots, x_n)\right),$$

that is, by convention,

$$\forall x_1 \ldots \forall x_n \, \exists y \, \forall x_0 \, (x_0 \in y \leftrightarrow \varphi(x_0, x_1, \ldots, x_n)).$$

Again, notice that only logical axioms have been used in this derivation. While the formulation of the T-biconditional involves closed terms like $\ulcorner \varphi \urcorner$ (which is composed of the successor function symbol S and individual constant $\overline{0}$)[7] and function symbols like $\dot{s}_1^1$, no non-logical axioms governing the behaviour of these symbols are assumed anywhere in the derivation. For example, it is nowhere assumed that $\dot{s}_1^1(\ulcorner \varphi \urcorner, t) =$

[7]Notice that in formulating the T-biconditionals, we have also assumed a Gödelcoding in the metalanguage in order to pick a name for each formula. However, for the derivation (and the results below) it is entirely irrelevant which name is actually chosen, and we could simply pick names at random, with the proviso that no two formulae get the same name assigned.

$\ulcorner\varphi(t)\urcorner$ or that $\overline{0} \neq S(\overline{0})$. However, what is important is that the function symbols $\dot{s}^n_n$ are primitive. If they were eliminated using defining arithmetical formulas, one could not ensure totality and functionality, and the derivation would fail.

The fact that uniform disquotation sentences behave like instances of naïve comprehension has some interesting consequences, as we will show next. First, unlike the ordinary (i.e., non-uniform) T-schema, the uniform T-schema leads to inconsistency over logic alone. As Gupta [9] has shown, a theory can be classical, contain names for its own expressions, obey the ordinary T-schema, and be consistent at the same time. Only when in addition self-referential liar-like statements can be formulated do we get inconsistency. In stark contrast, the uniform T-schema can lead to inconsistency even in the absence of self-reference. Second, restricted (and consistent) instances of the uniform T-schema, taken over logic alone, commit us to the existence of infinitely many objects and suffice to interpret certain amounts of arithmetic, so that the set of their consequences is essentially undecidable. In particular, typed uniform disquotation interprets R while uniform positive disquotation interprets Q. It goes without saying that the existence of infinitely many objects does not follow from the mere presence of infinitely many names in the language.

The usual way of formalizing the liar paradox in arithmetic is as follows. We want to obtain a fixed point of the predicate $\neg Tx$, that is, a term l such that $\mathsf{PA} \vdash l = \ulcorner\neg Tl\urcorner$. For, given such a term l, the T-schema yields

$$T\ulcorner\neg Tl\urcorner \leftrightarrow \neg Tl,$$

whence by substitutivity of identicals we obtain the contradiction

$$Tl \leftrightarrow \neg Tl$$

In order to produce such a liar sentence, we appeal to the diagonal lemma. Let $n := \#\neg T\dot{s}(x, x)$. Now observe that $s(n, n) = \#\neg T\dot{s}(\overline{n}, \overline{n})$, so PA proves $\dot{s}(\overline{n}, \overline{n}) = \ulcorner\neg T\dot{s}(\overline{n}, \overline{n})\urcorner$. So $\dot{s}(\overline{n}, \overline{n})$ is our desired term l. Notice that no contradiction follows from the (non-uniform) T-schema if we don't have the identity statement $\dot{s}(\overline{n}, \overline{n}) = \ulcorner\neg T\dot{s}(\overline{n}, \overline{n})\urcorner$ as a further premise. By contrast, the identity is not required in order to derive a contradiction from the uniform T-schema.

Proposition 27 *The scheme* $\forall x(T\ulcorner\varphi(\dot{x})\urcorner \leftrightarrow \varphi(x))$ *is inconsistent over classical first-order logic.*

Proof By a simple application of Russell's paradox. The uniform T-biconditional for the formula $\neg T\dot{s}(x, x)$ delivers, via the 'translation' provided at the beginning of this section, that $\exists y\forall x(x \in y \leftrightarrow x \notin x)$, which is logically inconsistent. □

Next, we show that the typed theory of uniform disquotation, taken over logic alone, interprets the arithmetical theory R introduced in Tarski et al. [35], which is essentially undecidable and has only infinite models. (Roughly, R is Q formulated with numerals. See also Monk [22, chap. 14].)

UTB is the theory, formulated in the language $\mathcal{L}_{\mathrm{T}}$, whose axioms comprise those of PAT and all sentences of the form:

$$\forall x_1 \ldots \forall x_n(T\ulcorner\varphi(\dot{x_1}, \ldots, \dot{x_n})\urcorner \leftrightarrow \varphi(x_1, \ldots, x_n)),$$

where φ is a T-free formula (i.e. a formula of $\mathcal{L}_{\mathrm{PA}}$). This theory is a conservative extension of PA. Let us denote by UTB^- the theory that results from UTB by dropping all non-logical axioms of PAT. Thus, this theory contains only the uniform T-biconditionals for the T-free formulae, plus first-order logic with identity. To be sure, while that theory is still formulated in the language $\mathcal{L}_{\mathrm{T}}$, none of the Peano axioms are assumed; e.g. we do not assume that $\overline{0} \neq S(\overline{0})$.

Proposition 28 UTB^- *relatively interprets* R.

Proof By a result of Visser [37], it is sufficient to show that UTB^- interprets Vaught's set theory VS. The theory VS is formulated in the language $\mathcal{L}_{\in}$ and contains the following axioms:

$$\exists y \, \forall x \neg x \in y$$

and

$$\forall y_1 \ldots \forall y_n \, \exists y \, \forall x \, (x \in y \leftrightarrow x = y_1 \vee \ldots \vee x = y_n)$$

for every $n \geq 1$. The first axiom gives us the empty set, while the second axiom gives us singletons, (unordered) pairs, triples, etc. It is shown in Visser [37] that VS interprets the theory R.

The formulae $x \neq x$ and $\varphi_n(x) := (x = y_1 \vee \ldots \vee x = y_n)$ (for every $n \geq 1$) are T-free. Thus UTB^- has as axioms

$$\forall x (T^{\ulcorner} \dot{x} \neq \dot{x}^{\urcorner} \leftrightarrow x \neq x)$$

and

$$\forall y_1 \ldots \forall y_n \forall x \, (T^{\ulcorner} \varphi_n(\dot{x}, \dot{y_1}, \ldots, \dot{y_n})^{\urcorner} \leftrightarrow x = y_1 \vee \ldots \vee x = y_n)$$

for every $n \geq 1$. Using the 'translation' provided at the beginning of this section, it is easily seen that the axioms of VS are interpretable in UTB^-. □

Notice that the above result does not imply that UTB^- can prove that $\overline{0} \neq S(\overline{0})$. Rather, UTB^- proves an interpretation of that claim, namely that

$$^{\ulcorner} x \neq x^{\urcorner} \neq {}^{\ulcorner} x = {}^{\ulcorner} x \neq x^{\urcorner \urcorner}$$

This is because VS interprets 0 as the empty set, that is the extension of the predicate $x \neq x$, and interprets 1 as the singleton of the empty set, that is the extension of the predicate $x = {}^{\ulcorner} x \neq x^{\urcorner}$.

UTB^- does not interpret Robinson arithmetic, Q, as an anonymous referee has demonstrated to me. On the other hand, we can show that the theory of *positive* uniform disquotation, over logic alone, does interpret Robinson arithmetic. Recall from Section 4 that PUTB is the theory, formulated in the language $\mathcal{L}_{\mathrm{T}}$, whose axioms comprise those of PAT and all sentences of the form:

$$\forall x_1 \ldots \forall x_n (T^{\ulcorner} \varphi(\dot{x_1}, \ldots, \dot{x_n})^{\urcorner} \leftrightarrow \varphi(x_1, \ldots, x_n)),$$

where φ is a T-positive $\mathcal{L}_{\mathrm{T}}$-formula. Here, a formula is called T-positive if no occurrence of the truth predicate is in the scope of an odd number of negation signs. Let us denote by PUTB^- the theory that results from PUTB by dropping all non-logical axioms of PAT.

Proposition 29 PUTB$^-$ *relatively interprets* Q.

Proof By a well known result, it suffices to show that PUTB$^-$ interprets adjunctive set theory, AS. (A proof of this result can be found in Visser [38], who also gives a short history of this result.) AS is the theory in $\mathcal{L}_\in$ and contains the following two axioms:

$$\exists y\, \forall x \neg x \in y$$

and

$$\forall z \forall w\, \exists y\, \forall x\, (x \in y \leftrightarrow x \in w \vee x = z)$$

The formulae $x \neq x$ and $(T\dot{s}(w, x) \vee x = z)$ are T-positive. Therefore, PUTB$^-$ contains the following two axioms:

$$\forall x (T\ulcorner \dot{x} \neq \dot{x}\urcorner \leftrightarrow x \neq x) \tag{11}$$

and

$$\forall z \forall w \forall x\, (T\ulcorner T\dot{s}(\dot{w}, \dot{x}) \vee \dot{x} = \dot{z}\urcorner \leftrightarrow T\dot{s}(w, x) \vee x = z) \tag{12}$$

Using the translation provided at the beginning of this section, it is seen that the axioms of AS follow from PUTB$^-$.

More precisely, Eq. 11 is shorthand for

$$\forall x (T\dot{s}(\ulcorner x \neq x\urcorner, x) \leftrightarrow x \neq x)$$

from which we deduce in pure logic

$$\exists y \forall x \neg T\dot{s}(y, x)$$

Let $\varphi := T\dot{s}(w, x) \vee x = z$. What Eq. 12 really says is:

$$\forall z \forall w \forall x\, \left(T\dot{s}\left(\dot{s}^2_2\left(\dot{s}^3_3(\ulcorner \varphi \urcorner, z), w\right), x\right) \leftrightarrow T\dot{s}(w, x) \vee x = z\right)$$

Instantiating the quantifiers $\forall z \forall w$ to z, w, we get:

$$\forall x\, \left(T\dot{s}\left(\dot{s}^2_2\left(\dot{s}^3_3(\ulcorner \varphi \urcorner, z), w\right), x\right) \leftrightarrow T\dot{s}(w, x) \vee x = z\right)$$

By existential weakening,

$$\exists y\, \forall x\, (T\dot{s}(y, x) \leftrightarrow T\dot{s}(w, x) \vee x = z)$$

Now we re-introduce the universal quantifers and get

$$\forall z \forall w\, \exists y\, \forall x\, (T\dot{s}(y, x) \leftrightarrow T\dot{s}(w, x) \vee x = z)$$

as desired. □

The results in this section show that a good amount of arithmetic is already encoded into the uniform T-biconditionals. As mentioned above, the use of the *primitive* function symbols $\dot{s}^n_n$ (for $n \geq 1$) in the formulation of the uniform T-schema plays an essential role (because we need to ensure totality and functionality), although the fact that there are infinitely many is of no relevance. Instead of the infinitely many symbols $\dot{s}^n_n$ we could use a single binary function symbol $\dot{h}$ in the formulation of the uniform T-schema. (On the intended interpretation, $\dot{h}$ denotes the function h such that

$h(m, n) = \#\varphi(\overline{n}/x_j)$ if $m = \#\varphi$ codes a formula with some free variables and x_j is the variable with the highest index in φ, and $h(m, n) = 0$ otherwise.)

7 Conclusions

In this paper we considered translations that map formulae of the language of second-order arithmetic to formulae of the language of truth. From a technical point of view, such translations are useful because they allow us to transform questions about truth into questions about second-order arithmetic and to exploit the extensive literature on the latter to answer questions about the former. We have seen several such applications in this paper. From a philosophical point of view, such translations indicate, I believe, that there is a close connection between truth and satisfaction on the one hand and second-order quantification on the other hand. The results in this paper, and other similar results in the literature, support Parsons' claim, mentioned in the introduction, that truth and satisfaction are a means to quantify into sentence respectively predicate position.

There are at least two directions into which the study of translations between second-order arithmetic and theories of truth ought to be further developed. First, our attention in this paper was restricted to ω-models and it might be interesting to generalize some of the results to non-standard models. Second, our attention in this paper was restricted to translations from second-order arithmetic into the language of truth. Of course, the reverse is also possible. It is well known that theories of truth can be interpreted in subsystems of second-order arithmetic too. As an example (see [11]), one may take the interpretation of typed compositional theories of truth (e.g. the systems of ramified truth RT_α) into systems of predicative comprehension (e.g. the systems of ramified analysis RA_α). This raises the question of the composition of such interpretations. In [24], Nicolai has shown that the mutual embeddings between typed truth and predicative comprehension cannot be lifted to stricter notions of equivalence such as bi-interpretability. For example, he shows that while for each $n \in \omega$, RT_n is an e-retract of RA_n, the converse does not hold. Further interesting work awaits us there.

Acknowledgments For their help in preparing this paper, I want to thank Timo Beringer, Catrin Campbell-Moore, Volker Halbach, Carlo Nicolai, Stanislav Speranski, and two anonymous referees, one of which spotted a flaw in an old proof of Proposition 14. A special thanks goes to Lavinia Picollo, in particular for her help with Proposition 21 .

References

1. Beringer, T., & Schindler, T. (2017). A graph-theoretic analysis of the semantic paradoxes. Under review.
2. Boolos, G., Burgess, J., & Jeffrey, R. (2003). *Computability and logic*, 4th edn. Cambridge: Cambridge University Press.

3. Burgess, J.P. (1986). The truth is never simple. *Journal of Symbolic Logic*, *51*, 663–681.
4. Cain, J., & Damnjanovic, Z. (1991). On the Weak Kleene scheme in Kripke's theory of truth. *Journal of Symbolic Logic*, *56*, 1452–1468.
5. Cantini, A. (1989). Notes on formal theories of truth. *Zeitschr. f. Math. Logik und Grundlagen d Math*, *35*, 97–130.
6. Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic*, *55*, 244–259.
7. Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, *56*, 1–49.
8. Fischer, M., Halbach, V., Kriener, J., & Stern, J. (2015). Axiomatizing semantic theories of truth? *Review of Symbolic Logic*, *8*, 257–278.
9. Gupta, A. (1982). Truth and paradox. *Journal of Philosphical Logic*, *11*, 1–60.
10. Halbach, V. (1995). Tarski hierarchies. *Erkenntnis*, *43*, 339–367.
11. Halbach, V. (1996). *Axiomatische Wahrheitstheorien Logica Nova*. Berlin: Akademie Verlag.
12. Halbach, V. (2000). Truth and reduction. *Erkenntnis*, *53*, 97–126.
13. Halbach, V. (2001). How innocent is deflationism? *Synthese*, *126*, 167–194.
14. Halbach, V. (2009). Reducing compositional to disquotational truth. *Review of Symbolic Logic*, *2*, 786–798.
15. Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
16. Herzberger, H. (1982). Notes on naive semantics. *Journal of Philosphical Logic*, *11*, 61–102.
17. Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In P. Cortois (Ed.), *The many problems of realism, vol. 3 of Studies in the general philosophy of science* (pp. 173-187). Tilburg University Press.
18. Horwich, P. (1998). *Truth*, 2nd edn. Oxford: Basil Blackwell.
19. Ketland, J. (1999). Deflationism and Tarski's paradise. *Mind*, *108*, 69–94.
20. Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, *72*, 690–716.
21. Leitgeb, H. (2005). What truth depends on. *Journal of Philosphical Logic*, *34*, 155–192.
22. Monk, D.J. (1976). *Mathematical logic*. New York: Springer.
23. Moschovakis, Y.N. (1974). *Elementary induction on abstract structures.* Dover Publications.
24. Nicolai, C. (2016). *Equivalences for truth predicates.* Review of Symbolic Logic. forthcoming.
25. Parsons, C. (1983). Ontology and mathematics. In Mathematics in philosophy, Cornell University Press, .
26. Parsons, C. (1983). Sets and classes. In *Mathematics in philosophy*, (pp. 209–220). Cornell University Press.
27. Pohlers, W. (2009). *Proof theory the first step into impredicativity*. Berlin Heidelberg: Springer.
28. Putnam, H., Boyd, R., & Hensel, G. (1969). A recursion theoretic characterization of the ramified analytical hierarchy. *Transactions of the American Mathematical Society*, *141*(/142), 37–62.
29. Schindler, T. (2014). Axioms for grounded truth. *Review of Symbolic Logic*, *7*, 73–83.
30. Schindler, T. (2015). A disquotational theory of truth as strong as Z_2^-. *Journal of Philosophical Logic*, *44*, 395–410.
31. Shapiro, S. (1998). Proof and truth: Through thick and thin. *Journal of Philosophy*, *95*, 493–521.
32. Shoenfield, J.R. (1967). *Mathematical logic.* Addison-Wesley.
33. Simpson, S.G. (2009). *Subsystems of second order arithmetic*, 2nd edn. Cambridge: Cambridge University Press.
34. Speranski, S. Notes on the computational aspects of Kripke's theory of truth. Studia Logica (2016), doi:10.1007/s1122501696948.
35. Tarski, A., Mostowski, A., & Robinson, R. (1953). Undecidable theories. North Holland, Amsterdam.
36. Visser, A. (1997). An overview of interpretability logic. In *Advances in modal logic*, (pp. 307–359). Stanford: CSLI Publications.
37. Visser, A. (2008). Pairs, sets and sequences in first-order theories. *Archive for Mathematical Logic*, *47*, 299–326.
38. Visser, A. (2009). Cardinal arithmetic in the style of Baron von Münchhausen. *Review of Symbolic Logic*, *2*, 570–589.
39. Welch, P. (2001). On Gupta-Belnap revision theories of truth, Kripkean fixed points, and the next stable set. *Bulletin of Symbolic Logic*, *7*, 345–360.
40. Welch, P. (2015). The complexity of the dependence operator. *Journal of Philosophical Logic*, *44*, 337–440.