AXIOMATIC TRUTH, SYNTAX AND METATHEORETIC REASONING

GRAHAM E. LEIGH
Faculty of Philosophy
University of Oxford
and
CARLO NICOLAI
Somerville College
University of Oxford

Abstract. Following recent developments in the literature on axiomatic theories of truth, we investigate an alternative to the widespread habit of formalizing the syntax of the object-language into the object-language itself. We first argue for the proposed revision, elaborating philosophical evidences in favor of it. Secondly, we present a general framework for axiomatic theories of truth with 'disentangled' theories of syntax. Different choices of the object theory O will be considered. Moreover, some strengthenings of these theories will be introduced: we will consider extending the theories by the addition of coding axioms or by extending the schemas of O, if present, to the entire vocabulary of our theory of truth. Finally, we touch on the philosophical consequences that the theories described can have on the debate about the metaphysical status of the truth predicate and on the formalization of our informal metatheoretic reasoning.

§1. Truth, Syntax and Object theory. In the customary way of constructing axiomatic theories of truth, starting with an object theory O—usually a formal system capturing some portion of arithmetic—the language $\mathscr{L}_O$ is expanded with a unary truth predicate T and O is extended with truth-theoretic axioms. Under a suitable coding for $\mathscr{L}_O$-expressions, the truth predicate then applies to names of sentences of $\mathscr{L}_O$ in the form of terms of $\mathscr{L}_O$ itself. Therefore primitive recursive functions and relations concerning strings of symbols of the language of O are represented in $\mathscr{L}_O$ in the guise of expressions naturally interpreted over the mathematical domain of O.

In the wake of recent works of Richard Heck and Volker Halbach, this paper stands as a contribution towards a possible alternative to this common habit. In the rest of the present section we try to motivate the proposed reassessment by highlighting some unresolved issues traceable in the usual construction.[1]

Received: November 28, 2012.

[1] Throughout the paper we will refer on occasion to Heck's unpublished manuscript (Heck 2009). Several ideas, together with some of the motivational remarks that are investigated in this work are already present there. The manuscript consists of two parts: in the first Heck considers disquotational and compositional axiomatizations of truth over sequential theories constructed in the usual way—i.e. in which the syntax is contained in the object theory—and investigates the differences between the cases in which principles of truth are added to finitely axiomatized or schematically axiomatized base theories by employing Solovay's technique of shortening of cuts. In the second part, Heck investigates theories of truth with 'disentangled' syntax. He works over

doi:10.1017/S1755020313000233

1.1. Informal metatheoretic reasoning. Volker Halbach in (Halbach 2011) expresses manifest uneasiness with the praxis of identifying the formal system S in which we talk about the syntax of the language $\mathscr{L}_O$ of a system O and O itself. The typical situation, when axiomatic truth is concerned, regards the case in which O is a fragment of arithmetic which contains S as a subtheory:

> Identifying numbers and expressions is a notational simplification at best, but in informal metatheoretic discussion the theory of syntax and the theory of natural numbers should be kept separate: expressions are not numbers. (Halbach, 2011, p. 316)

In other words, in 'informal metatheoretic discussion' it is always possible for the working logician to look at the content of some expression e and determine whether it describes syntactic properties of the language of O or whether it refers instead to the subject-matter of O. Suppose O is a theory of arithmetic and consider for instance the operation (which we shall label $\check{\neg}$) that given an expression of $\mathscr{L}_O$ prepends to it the negation symbol '$\neg$'. This operation belongs to our informal metatheory and in the customary way of formalizing the metatheory is expressed by a function $\hat{\neg}$ which is then represented in $\mathscr{L}_O$ by a term $\dot{\neg}(x)$ for which $\dot{\neg}\ulcorner\varphi\urcorner = \ulcorner\hat{\neg}\varphi\urcorner$ holds in every model of O, with the notation $\ulcorner.\urcorner$ representing some fixed Gödel coding of expressions from $\mathscr{L}_O$. From the perspective of the theory O therefore, an expression concerning the syntax of $\mathscr{L}_O$ is indistinguishable from one concerning its subject-matter. In contrast, in the language of the formal metatheory all statements are syntactic. Thus, from the point of view of our informal metatheory, we clearly distinguish between $\check{\neg}$, $\hat{\neg}$ and $\dot{\neg}$.

The theories that will be proposed in what follows, especially in section 3.4, will try to capture this all-present interplay between formal and informal components of our metamathematical reflection.

1.2. Generality. A formal theory of truth, from Tarski onwards,[2] has to be applicable, *sine contradictione*, as widely as possible. For instance it has been argued in Horsten (2011) that the addition of an axiomatized truth predicate to epistemological or metaphysical theories could in some cases shed light on the explanatory power of the notion of truth.[3] For instance, let us define a mereological theory as a first-order theory formulated in the language of first-order logic augmented with a binary predicate $\circ$ for 'overlapping' or alternatively $\sqsubseteq$ for 'part-of', and which is strong enough to contain the so-called *calculus of individuals* (*CI*).[4] We might want to expand the language of our mereological theory M

---

arithmetical object theories and employs restricted versions of syntactic induction, so the setting is slightly less general then the one presented below. It should also be noted that Heck attributes to Albert Visser several results contained in his paper.

[2] Cfr. §1.3.

[3] Horsten considers the case of Fitch's argument against weak verificationism and shows that the truth predicate, characterized by more than just the T-sentences, does play a substantial role in the argument. Cfr. Horsten (2011, pp. 86–91).

[4] For instance, as defined in Niebergall (2011), *CI* is the first-order theory in $\mathscr{L}_\circ$, where $\circ$ is a primitive predicate for overlapping, whose nonlogical axioms are (with a predicate $\sqsubseteq$ for 'part-of' relation already definitionally introduced):

$$\forall x, y(x \circ y \leftrightarrow \exists z(z \sqsubseteq x \wedge z \sqsubseteq y)) \qquad (O)$$

$$\forall x, y \exists z \forall u(u \circ z \leftrightarrow u \circ x \vee u \circ y) \qquad (SUM)$$

$$\forall x(\neg\forall v(v \circ x \rightarrow \exists z \forall v(v \sqsubseteq z \leftrightarrow \neg v \circ x))) \qquad (NEG)$$

with a unary truth predicate and axiomatize it according to our needs. Now there is no consistent extension M' of a mereological theory so-defined such that Robinson's arithmetic Q is interpretable in it.[5] Assuming that the strength of Q is the minimal requirement to develop a satisfactory theory of syntax up to diagonalization, there seems to be no way for M' to develop its own syntax, and a separate axiomatization of the syntax for $\mathscr{L}_\circ$ or $\mathscr{L}_\sqsubseteq$ is particularly welcome.

1.3. Tarski. Surprisingly enough, Tarski's original presentation already displays several of the aspects of metamathematical reflection that Halbach's thesis emphasizes. In particular, Tarski characterized the metalanguage as containing

> ... three groups of expressions: (1) expressions of a general logical kind; (2) expressions having the same meaning as all the constants of the language to be discussed [...]; (3) expressions of the structural-descriptive type which denote single signs and expressions of the language considered, whole classes and sequences of such expressions or, finally, the relations existing between them. (Tarski, 1936, §4, pp. 210–11)

The description of the metalanguage is then completed by specifying the axiom system for it, compounded by

> ... the *general* logical axioms which suffice for a sufficiently comprehensive system of mathematical logic, and the *specific axioms of the metalanguage* which describe certain elementary properties of the above structural-descriptive concepts consistent with our intuitions. (*Ibid*., §2, p. 173)

According to Heck the construction outlined in the passages above seems to be that Tarski's project of providing metamathematics with a consistent notion of truth does not depend on the expressions of the structural-descriptive type—group (3) in the quotation—being part of expressions of category (2). We agree on this interpretation.

The reasons behind a separate axiomatization of the syntax, in Tarski (1936), seem to be deeply tied with the dimension of generality he wanted to give to his definition. The following passage calls to mind the content of the previous subsection:

> In contrast to natural languages, the formalized languages do not have the universality which was discussed at the end of the preceding section. In particular, most of these languages possess no terms belonging to the theory of language, i.e. no expressions which denote signs and expressions of the same or another language or which describe the structural connexions between them (such expressions I call—for a lack of a better term—*structural-descriptive*). For this reason, when we investigate the language of a formalized deductive science, we must always distinguish clearly between the language *about* which we speak, as well as between the science which is the object of our investigation and the science in which the investigation is carried out. (*ibid.*, §2, p. 167)

---

[5] *Ibid.*, p. 290.

This famed passage seems also to indicate that when the language $\mathscr{L}_O$ of our object theory O *does* indeed contain expressions 'which denote signs and expressions of the same or another language or which describe the structural connexions between them', it would be natural for the object theory to contain the syntactic portion of the metatheory. This interpretation appears to be supported also by Tarski's more general description of the 'semantic conception of truth' contained in Tarski (1944, §9), in which the above-mentioned condition on the object language is explicitly made.

Tarski's picture of the metatheory thus embraces the possibility of a setting in which the syntactic component of the metatheory itself is not included in the object theory. This seems to leave room for the investigation of a setting in which syntax and mathematics are clearly distinguished.[6]

Tarski aimed at a definition of truth in the metatheory, whereas in our approach truth will be taken to be a primitive concept characterized by suitable axioms. One might then wonder whether this difference can impinge on the motivations behind the project carried out here.[7] As it will be clear below, one of the goals of an axiomatization of Tarski's original picture of the metatheory is to isolate the different patterns of reasoning proper of our meatmathematical reflection which are coarsely identified in the usual construction. Therefore a setting in which the notion of truth is described by axioms renders the evaluation of the syntactic, truth theoretic and mathematical components of our metatheoretic reasoning easier. In a definitional approach, as in Tarski's original one, truth is reduced to other mathematical notions such as higher-order quantification: the impact of the semantic constituents of the metatheory would thus be dissolved in a richer mathematical structure.

As a matter of fact, the proposal that we are advocating will not be fully completed in this paper. Unlike Tarski, we will favor the arithmetization techniques, as the presentation of a brand new theory of expressions for truth-theoretic purposes would require the length of another paper. However, following Heck (2009) and Halbach (2011), we will distinguish between a realm of syntactic objects and the domain of discourse of our object theory. Therefore expressions and numbers in our system are not the same kind of objects. We provide a concrete, sufficiently general, and hopefully accessible account of how axiomatic theories of truth with a detached theory of syntax can be constructed, of some of their properties and of the some ways in which they can be strengthened. We stick to the case of a classical axiomatization of the notion of truth for a first-order object theory O, and defer the treatment of the self-referential setting to future work. Before that, however, we consider further motivations, besides the one already singled out, for endorsing our proposed revision.

1.4. 'Syntactic' and 'mathematical' schemas. We assume some familiarity with the axiomatization of classical Tarskian truth as presented for instance in Feferman (1991) and Halbach (2011). We begin with a sufficiently powerful object theory O formulated in the language $\mathscr{L}$ of arithmetic,[8] expand its language with a unary truth predicate, allowed to appear into the schemas of the object theory, and add compositional axioms of the form

$$\forall x \forall y (Sent_{\mathscr{L}_O}(x \mathbin{\dot{\wedge}} y) \rightarrow (Tx \mathbin{\dot{\wedge}} y \leftrightarrow Tx \wedge Ty)) \qquad (CT\wedge)$$

[6] For further insights on the exegesis of Tarski's work, also in relation with Gödel's theorems, see Nicolai (forthcoming).

[7] We would like to thank an anonymous referee for suggesting some clarificatory remarks on this matter.

[8] Here by sufficiently powerful we mean the capability of developing its own syntax.

for each connective and quantifier of $\mathscr{L}$, where the $\mathscr{L}$-expression $x \mathbin{\dot{\wedge}} y$ stands for the term of $\mathscr{L}$ representing the operation of applying the conjunction symbol between the $\mathscr{L}$-sentences represented by x and y.[9] Following Halbach (2011), we call the resulting theory *CT*, for 'compositional truth' when the object theory O is Peano Arithmetic, specifying otherwise the object theory (writing $CT[O]$). We will occasionally refer to the theory $CT\restriction$ as *CT* with the schema of induction restricted to formulae of $\mathscr{L}$.

It is well-known that for various choices of the object theory O, for O finitely and schematically axiomatized,[10] the theory of truth $CT[O]$ is strong enough to prove the so-called *global reflection principle* for O.

PROPOSITION 1.1. *For suitable O,* $CT[O] \vdash \forall x(Sent_{\mathscr{L}_O}(x) \wedge Bew_O(x) \rightarrow Tx)$.

As immediate corollaries of Proposition 1.1, $CT[O]$ will prove the consistency statement for O. By Gödel's second incompleteness theorem, $CT[O]$ is thus nonconservative over O. In Heck (2009), it is shown that the full deductive strength of *CT* is not needed to prove the global reflection principle for *PA*. In particular:

PROPOSITION 1.2 (HECK). $CT[I\Sigma_1] \vdash \forall x(Sent_{\mathscr{L}}(x) \wedge Bew_{PA}(x) \rightarrow Tx)$

Proposition 1.2 is particularly striking: if we start with $I\Sigma_1$, expand its language with a predicate expressing truth and extend the theory with compositional axioms, we expect to gain some expressive or deductive power over our object theory. In this case however, we do not only obtain *some* increase in deductive power, but rather obtain a theory whose deductive strength is sufficient to prove the consistency of Peano arithmetic. $CT[I\Sigma_1]$, *a fortiori*, would thus prove the consistency of the theory $I\Sigma_i$ for each i. What is crucial for our purposes is that Proposition 1.2 is only provable because the theory of syntax for $\mathscr{L}$ is considered to be part of the object theory itself. It is proved by formal induction on the length of the derivations in the object theory, which crucially employs the extension of the schemas in the object theory to formulas containing the truth predicate. In particular, we need the following instance of the induction of $CT[I\Sigma_1]$:

$$\forall x\big(\forall y < x(Proof_{PA}(y) \rightarrow T(end(y))) \rightarrow (Proof_{PA}(x) \rightarrow T(end(x))\big) \rightarrow \forall x(Proof_{PA}(x) \rightarrow T(end(x))) \tag{1}$$

where $Proof_{PA}(z)$ is some standard primitive recursive predicate expressing that 'z encodes a proof in a sentential calculus for *PA*' and $end(z)$ is the primitive recursive function that outputs the (code of the) final formula of the proof encoded in z. However, in order to obtain the premise of (1), in particular the claim that the universal closure of each axiom of *PA* is true, we need another instance of the extended induction, namely,

$$Sent_{\mathscr{L}}(subn(z, \bar{0})) \rightarrow \big(T(subn(z, \bar{0})) \wedge \forall x(T(subn(z, x)) \rightarrow T(subn(z, Sx))) \rightarrow \forall x T(subn(z, x))\big) \tag{2}$$

where $subn(z, x)$ is the numeral substitution function, representing within $\mathscr{L}_O$ the operation of substituting the x-th numeral for the first free variable occuring in the formula represented by z. Now (1) and (2) are, in a sense, very different. (1) is carried out on syntactic objects, derivations within *PA*, whereas in (2) the induction is carried out over all substitutional instances of the formula encoded by z. It is worth noting the ambiguity that

[9] Likewise for other connectives and quantifiers in $\mathscr{L}$.

[10] See the beginning of §2, for a precise definition of 'schematic theory'.

arises once again: technically speaking, both are instances of the induction of *PA* in the language $\mathscr{L} \cup \{T\}$, but it is clear that for our metamathematical concerns they are part of quite different kinds of argument. It is only because of the identification between syntax and object theory that the proof is possible in the case of *CT*. In the concluding section we will discuss this duplex nature of the induction schema of *PA* in the light of the so-called 'conservativeness argument' against deflationism raised in Horsten (1995), Shapiro (1998), and Ketland (1999). As we shall see, the difference between 'syntactic' and 'mathematical' versions of the induction schema of *PA* is particularly relevant for that debate.

In another sense, however, Theorem 3.4 is not surprising at all, as the truth predicate allowed into the Σ^0_1-induction scheme determines a 'collapse' of the arithmetical hierarchy, as formulas of any complexity are treated as atomic. This fact becomes more evident if we consider the case of set theory. Abusing a bit of the notation just introduced we call $CT[KP\omega]$ the theory obtained from Kripke-Platek set theory plus the axiom of infinity by expanding $\mathscr{L}_\in$ with constants c_x for each element x of the the intended model of set theory we are referring to and a unary truth predicate which is allowed to appear in the Δ_0-collection and Δ_0-separation axiom schemas, and by adding compositional axioms such as ($CT\wedge$) to it.[11] It is relatively easy to see that:

PROPOSITION 1.3. $CT[KP\omega] \vdash \forall x(Sent_{\mathscr{L}_\in}(x) \wedge Bew_{\mathrm{ZF}}(x) \rightarrow Tx)$

Proof. The proof is by ω-induction on the length of the derivation in *ZF*. The truth of all the finitely many axioms of *ZF* is obtained from the provability in $CT[KP\omega]$ of the Tarskian uniform disquotation scheme for $\mathscr{L}_\in$, i.e. T-biconditionals for formulas containing free variables. The truth of the schemas of *ZF* at each substitutional instance is provable in $CT[KP\omega]$ in the manner of (2), that is via suitable instances of the extended schemas of Δ_1-separation and Σ_1-replacement (with the truth predicate) that are provable in $CT[KP\omega]$.[12] By the same token, we have ω-induction extended to contain the truth predicate, which is needed for the induction step in the same way as (1). □

At any rate, the proof is again only possible because the schemas of our truth theory have instances that entail the truth of the schemas of the object theory at each substitutional instance. Let us reinforce this point: having ω-induction for the language of the base theory extended to semantic vocabulary is a property of our metatheory. We want in fact to allow ourselves with syntactic arguments involving the notion of truth. However, we only get the necessary ω-induction as a *theorem* of the object theory because the metatheory is assumed to be part of $KP\omega$.

The theories of truth with disentangled syntax will solve the ambiguities that led to Propositions 1.2 and 1.3. In our informal reasoning we in fact implicitly distinguish between a sort of variables ranging over syntactic objects, and a sort of variables ranging over the domain of discourse of our object theory.

§2. The theory *CTD*[*O*]. In the present section we introduce the three-sorted theory called $CTD[O]$, standing for 'compositional truth with disentangled syntax' for the object

---

[11] For a more detailed presentation of Tarskian truth over set theory, see Fujimoto (2012, pp.13ff). $KP\omega$ is treated extensively in Barwise (1975). We employ $KP\omega$ as object theory in our example as we know that it is a safe place in which we can develop the syntax for $\mathscr{L}_\in$ and prove that many useful syntactic notions are Δ_1-definable.

[12] See for instance Barwise (1975, p.17).

theory O. $CTD[O]$ is a generalization of the compositional system already presented in Heck (2009) for the case in which the base theory is an arithmetical theory. We first fix some notation. T is *finitely axiomatizable* if there is an axiomatization T' of T whose set of nonlogical axioms is finite. More importantly, we call T *schematically axiomatizable* (in the sense of Feferman (1991) and Lavine (1999)) if there is an axiomatization of T involving all substitutional instances of a finite set of schemas $\Phi(X)$ in which X is a free predicate variable. Examples of schematically axiomatizable theories are *PA*, in which each instantiation of the induction axiom can be represented by the single schema $X(\bar{0}) \wedge \forall x(X(x) \rightarrow X(Sx)) \rightarrow \forall x\, X(x)$, and *ZF*, where the separation and replacement axioms are similarly represented. In what follows, we will then call a theory *finitely axiomatized* or *schematically axiomatized* referring to a specific axiomatization of it.

2.1. Language and axioms of CTD[O]. The language $\mathfrak{L}$ of $CTD[O]$ is a first-order, three-sorted language. We start with a set of sorts $I = \{o, s, sq\}$, standing respectively for 'mathematical objects', 'syntax', and 'sequences'. I is intended as a set of expressions whose elements are employed to label variables of $\mathfrak{L}$. The first sort o labels variables ranging over the domain of the object theory O, the sort s labels variables ranging over the domain of the theory of syntax, while the third sort, sq, labels variables ranging over a third domain of variable assignments, 'mixed' objects of both syntactic and mathematical nature as they associate to syntactic objects elements of the domain of discourse of O. For clarity we will employ elements of I also to label nonlogical constants of $\mathfrak{L}$ when necessary.

Among the logical symbols of $\mathfrak{L}$, besides $\neg, \wedge, \forall$, we include a disjoint family of non empty sets

$$V = V_o \sqcup V_s \sqcup V_{sq}$$

of countably many variables of each sort. In what follows, we denote with $v_1, v_2, v_3, \ldots$, and occasionally with $x, y, z, \ldots$ variables in V_o, with $i, j, k, m, n, \ldots$ variables ranging over the domain of our theory of syntax S, and with $a, b, c, \ldots$, variables in V_{sq}. Each sort of variables can be expanded with indexes as necessary.

For the sake of the present paper, we consider our theory of syntax S to be formulated in a copy of the language of arithmetic. In particular, we set $\mathscr{L}_S = \{0^s, S^s, +^s, \times^s\}$. Besides the nonlogical constants of $\mathscr{L}_O$ and $\mathscr{L}_S$, $\mathfrak{L}$ also contains nonlogical symbols of a 'mixed' nature. In particular, $\mathfrak{L}$ comprehends two function symbols $.(.)$ and D, both of type $\langle sq, s\rangle \rightarrow o$ which apply to a sequence and a syntactic object and return a mathematical object, and a binary 'satisfaction' predicate *Sat* of type $\langle sq, s\rangle$.[13]

The axioms of $CTD[O]$ are displayed in Table 1.

$CTD[O]$ contains, as its subtheories, O itself; the basic axioms of *PA* as axioms for the syntax S; a simple theory of sequences *SQ* formulated in the language $\mathfrak{L}$ and consisting of one single axiom (axiom III.(*SQ*) in Table 1) which suffices to establish that there is at least one nonempty sequence and that given a variable assignment a we can always find a sequence b which differs from a only in what it assigns to one single variable;[14] and crucially the axiom of mathematical induction Ind^s for syntactic variables open to

[13] Here $.(.)$ has the same role that Gödel's β-function plays in the usual setting, and D is just a generalization to an evaluation function. The function D is dispensable if $\mathscr{L}_O$ is a relational language.

[14] This axiom, contained in a slightly different form in Heck (2009), is essentially due to Craig & Vaught (1958).

Table 1. *Axioms of CTD*$[O]$

(I)	Axioms of O	
(II)	Axioms of S (basic axioms of *PA*)	
(III)	Axiom for sequences of variable assignments:	
	(SQ)	$\forall a \forall x \forall j \exists b (\forall i (i \neq j \rightarrow a(i) = b(i)) \wedge b(j) = x)$
(IV)	Axioms for denotation and satisfaction:	
	$(CTDv)$	$\forall a \forall k (D(a, \dot{v}_k) = a(k))$
	$(CTDc)$	$\forall a \forall k (D(a, \dot{c}_k) = c_k)$
	$(CTDf)$	$\forall a \forall \ulcorner t_1 \urcorner, ..., \forall \ulcorner t_n \urcorner \big(D(a, \dot{f}^{\ulcorner} t_1 \urcorner, \ldots, \ulcorner t_n \urcorner) = f(D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))\big)$ for each n-ary function symbol f of $\mathscr{L}_O$.
	$(CTDat)$	$\forall a \forall \ulcorner t_1 \urcorner, ..., \forall \ulcorner t_n \urcorner \big(Sat(a, \dot{R}^{\ulcorner} t_1 \urcorner, \ldots, \ulcorner t_n \urcorner) \leftrightarrow R(D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))\big)$ for each n-ary relation symbol R in $\mathscr{L}_O$.
	$(CTD\neg)$	$\forall a \forall \ulcorner \varphi \urcorner \big(Sat(a, \dot{\neg} \ulcorner \varphi \urcorner) \leftrightarrow \neg Sat(a, \ulcorner \varphi \urcorner)\big)$
	$(CTD\wedge)$	$\forall a \forall \ulcorner \varphi \urcorner \forall \ulcorner \psi \urcorner \big(Sat(a, \ulcorner \varphi \urcorner \dot{\wedge} \ulcorner \psi \urcorner) \leftrightarrow Sat(a, \ulcorner \varphi \urcorner) \wedge Sat(a, \ulcorner \psi \urcorner)\big)$
	$(CTD\forall)$	$\forall a \forall \ulcorner \varphi \urcorner \forall i \big(Sat(a, \dot{\forall} v_i \ulcorner \varphi \urcorner) \leftrightarrow \forall b (\forall j (j \neq i \rightarrow a(j) = b(j)) \rightarrow Sat(b, \ulcorner \varphi \urcorner))\big)$
(V)	Syntactic induction:	
	(Ind^s)	$\varphi(0^s) \wedge \forall k (\varphi(k) \rightarrow \varphi(S^s k)) \rightarrow \forall k \varphi(k)$
	where k is an individual variable of $\mathscr{L}_S$ and φ is a formula of $\mathfrak{L}$	

arbitrary formulas of $\mathfrak{L}$. The choice of arithmetic as the theory of syntax allows us to obtain in one step the formalization of the syntax of $\mathscr{L}_O$ within $CTD[O]$.[15] In particular, under a canonical Gödel numbering of $\mathscr{L}_O$-expressions, we assign to the primitive symbols v, c, f, R, $\neg$, $\wedge$, $\forall$ of $\mathscr{L}_O$ provably distinct terms of $\mathscr{L}_S$, denoted by:

$$\dot{v}, \dot{c}, \dot{f}, \dot{R}, \dot{\neg}, \dot{\wedge}, \dot{\forall}$$

Then in $\mathscr{L}_S$, in which a primitive recursive pairing function is available (as well as its inverses) such that $(k, l) = n \rightarrow k, l < n$, we set

$$\dot{v}_k = (\dot{v}, k); \qquad \dot{c}_k = (\dot{c}, k);$$
$$\dot{f}_k^{\,l} = (\dot{f}, (k, l)); \qquad \dot{R}_k^{\,l} = (\dot{R}, (k, l)).$$

In order to generate syntactic codes for complex expressions we set:

$$\dot{f}_k^{\,l} k_1, \ldots, k_l = (\dot{f}_k^{\,l}, k_1, \ldots, k_l); \qquad \dot{R}_k^{\,l} k_1, \ldots, k_l = (\dot{R}_k^{\,l}, k_1, \ldots, k_l);$$
$$(\dot{\neg} k) = (\dot{\neg}, k); \qquad (k \,\dot{\wedge}\, l) = (\dot{\wedge}, (k, l));$$
$$(\dot{\forall} v_m) l = (\dot{\forall}, (\dot{v}_m, l)),$$

where each k_i with $1 \leq i \leq l$ codes an element of the set $Term_{\mathscr{L}_O}$ of $\mathscr{L}_O$ and k, l code elements of the set $Fml_{\mathscr{L}_O}$ of formulas of $\mathscr{L}_O$. This allows us to obtain for every expression e of the language $\mathscr{L}_O$ a term $\ulcorner e \urcorner$ of $\mathscr{L}_O$ itself representing the Gödel number $\sharp e$ of e. Therefore we can define formulas $Var^s_{\mathscr{L}_O}(k)$, $Const^s_{\mathscr{L}_O}(k)$ binumerating in S the set $Var_{\mathscr{L}_O}$ of variables and the set $Const_{\mathscr{L}_O}$ of constants of $\mathscr{L}_O$ respectively, and the

[15] See for instance Smoryński (1977).

formula $Term^s_{\mathscr{L}_O}(\ulcorner t \urcorner)$ binumerating in S the set $Term_{\mathscr{L}_O}$. Similarly we obtain formulas $AtFml^s_{\mathscr{L}_O}(\ulcorner \varphi \urcorner)$, $Fml^s_{\mathscr{L}_O}(\ulcorner \varphi \urcorner)$, $Sent^s_{\mathscr{L}_O}(\ulcorner \sigma \urcorner)$ binumerating in S respectively the sets $AtFml_{\mathscr{L}_O}$ of atomic formulas, $Fml_{\mathscr{L}_O}$ and the set $Sent_{\mathscr{L}_O}$ of sentences of $\mathscr{L}_O$. Formulas of $\mathscr{L}_S$ concerning syntactic categories of $\mathscr{L}_O$, for a matter of clarity, will be indexed by the superscript s.

In what follows, we will employ some notational shortcuts: as we already did in the last paragraph, we will employ $r, s, t, \ldots$ as ranging over elements of the set $Term_{\mathscr{L}_O}$, $\varphi, \psi, \zeta, \theta, \ldots$ for elements of $Fml_{\mathscr{L}_O}$, $\rho, \sigma, \tau, \ldots$ for elements of $Sent_{\mathscr{L}_O}$. Moreover, we will write, with $\varphi \in Fml_{\mathscr{L}_O}$, $\forall v_k \varphi$ instead of $(\forall v_k)\ulcorner \varphi \urcorner$, and we will quantify directly over formulas between Gödel corners so that

$$\forall \ulcorner \varphi \urcorner \text{ is intended to be short for } \forall k(Fml^s_{\mathscr{L}_O}(k) \to \ldots) \text{ and}$$
$$\forall \ulcorner t \urcorner \text{ is short for } \forall k(Term^s_{\mathscr{L}_O}(k) \to \ldots).$$

The substitution function expressing the result of formally substituting all free occurrences of the variable v_k in the formula φ with the code, within $\mathscr{L}_S$, of the term $t \in Term_{\mathscr{L}_O}$ can be defined in such a way that

$$S \vdash sub^s(\ulcorner \varphi \urcorner, k, \ulcorner t \urcorner) = \ulcorner \varphi(t) \urcorner$$

We will write $\varphi[t/v_k]$ for $sub^s(\ulcorner \varphi \urcorner, k, \ulcorner t \urcorner)$, when it is clear that we are working in S. We denote by $\varphi[t_1/v_1, \ldots, t_n/v_n]$ the result of simultaneously substituting the free variables $v_1, \ldots, v_n$ with the codes of $t_1, \ldots, t_n$ in the formula φ.

In §1.4 we emphasized the importance of the extended induction of CT (and of $CT[I\Sigma_1]$) in establishing the claim 'all instances of the induction axiom scheme of *PA* are true'. There we remarked that the proofs of Proposition 1.1 and Proposition 1.2 crucially relied on a double use of the extended induction of $CT[I\Sigma_1]$: one syntactic and one mathematical. Axiom Ind^s is thus designed to preserve that syntactic role, while blocking the mathematical one as displayed in (2). As a consequence, once the theory of syntax is disentangled from the underlying mathematics, we are able to keep safe our capability of formulating syntactic arguments involving the notion of truth, but also to avoid any peculiar interaction between syntax and object theory leading to objectionable consequences.

§3. The 'strength' of $CTD[O]$.

3.1. 'Syntactic' Reflection. As first result on $CTD[O]$, we prove the so-called *uniform disquotation scheme*. Tarski's disquotation scheme, an essential component of his Convention T, was originally presented in the form

$$\big(\forall a\, Sat(a, \ulcorner \sigma \urcorner)\big) \leftrightarrow \sigma \tag{3}$$

for each σ in the class of sentences of $\mathscr{L}_O$. The uniform disquotation scheme is instead defined for all formulas $\varphi(v_1, \ldots, v_n)$ of $\mathscr{L}_O$ with at most $v_1, \ldots, v_n$ free. We define an intermediate system:

$$CTD[O]\restriction := CTD[O] - Ind^s.$$

LEMMA 3.1. *For each formula φ of $\mathscr{L}_O$ with at most $v_1, \ldots, v_n$ free:*

$$CTD[O]\restriction \vdash \forall a \forall \ulcorner t_1 \urcorner, \ldots, \ulcorner t_n \urcorner \big(Sat(a, \varphi[t_1/v_1, \ldots, t_n/v_n]) \leftrightarrow \varphi(D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))\big)$$

Proof. The proof is by metatheoretic induction on the complexity of φ. A sublemma is required: if r is a term of $\mathscr{L}_O$ with at most $v_1, \ldots, v_n$ free then

$$\forall \ulcorner t_1 \urcorner, \ldots, \ulcorner t_n \urcorner \forall a \big(D(a, \ulcorner r(t_1, \ldots, t_n) \urcorner) = r(D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner)) \big) \tag{4}$$

is provable in $CTD[O]\restriction$. But this is straightforward to show by induction on the complexity of r given the axioms for D. The case in which φ is atomic, as well as the cases of propositional connectives, then follow easily from the axioms of $CTD[O]$. In the crucial case of the universal quantifier, in which $\varphi(v_1, \ldots, v_n)$ is $\forall v_j \psi(v_j, v_1, \ldots, v_n)$ we have to show, with $j > n$ that

$$\forall \ulcorner t_1 \urcorner, \ldots, \ulcorner t_n \urcorner \forall a \big(Sat(a, \forall v_j \psi[t_1/v_1, \ldots, t_n/v_n]) \leftrightarrow \forall v_j \psi(v_j, D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner)) \big).$$

For the left-to-right direction, from the axiom $CTD\forall$ we get $Sat(b, \ulcorner \psi(v_j, t_1, \ldots, t_n) \urcorner)$ for all sequences b which differ from a for what they assign to the j^{th} variable. Now by the induction hypothesis this yields $\psi(b(j), D(b, \ulcorner t_1 \urcorner), \ldots, D(b, \ulcorner t_n \urcorner))$. Since the value of $b(j)$ is arbitrary, $\forall v_j \psi(v_j, D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))$ is obtained by axiom SQ. For the converse direction, assume $\psi(v_j, D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))$. Then $\psi(b(j), D(b, \ulcorner t_1 \urcorner), \ldots, D(b, \ulcorner t_n \urcorner))$ holds for every b such that $b(i) = a(i)$ if $i \neq j$ by SQ. The induction hypothesis now applies, yielding $Sat(b, \ulcorner \psi(v_j, t_1, \ldots, t_n) \urcorner)$ from which $Sat(a, \ulcorner \forall v_j \psi(v_j, t_1, \ldots, t_n) \urcorner)$ results. □

Now if the theory O is finitely axiomatized, the theory of truth $CTD[O]$ behaves exactly like CT[O]. In fact we have the following:

PROPOSITION 3.2. *Let O be a finitely axiomatized theory.*

$$CTD[O] \vdash \forall a \forall \ulcorner \varphi \urcorner (Bew^s_O(\ulcorner \varphi \urcorner) \rightarrow Sat(a, \ulcorner \varphi \urcorner)).$$

The proof is by induction on the length of the derivations in O. The base step follows immediately from Lemma 3.1. For logical axiom schemas, it is always assumed that although we have a strict separation between syntax and mathematics, we do not want any separation between the logics behind O and $CTD[O]$.[16] The truth of all the instances of the logical axiom schemas is then obtained in a manner analogous to proving the truth of all instances of induction as presented in Proposition 1.2. For the induction step, as for Proposition 1.1 and Proposition 1.2, we need a syntactic inductive argument involving the notion of truth. This is exactly what axiom the Ind^s enables. In particular, the claim follows by Ind^s applied to the formula $\forall y < x(Proof^s_{PA}(y) \rightarrow \forall a(Sat(a, end^s(y))))$, that is the formulation of (1) in the language $\mathscr{L}_S$.

We denote the 'syntactic' global reflection principle for O with GRP^s_O. By employing the same notation, Proposition 3.2 entails:

COROLLARY 3.3. $CTD[O] \vdash Con^s_O$ *if O is finitely axiomatized.*

Proof. As an instance of GRP^s_O we have

$$Bew^s_O(\ulcorner \xi \urcorner) \rightarrow Sat(a, \ulcorner \xi \urcorner),$$

where ξ is a formula expressing an absurdity in O. By Lemma 3.1 we obtain $Bew^s_O(\ulcorner \xi \urcorner) \rightarrow \xi$. But also $\neg\xi$. Therefore $\neg Bew^s_O(\ulcorner \xi \urcorner)$, i.e. Con^s_O. □

[16] We judge this fact philosophically harmless, as we want logical validity to be transferred from our object theory to our theory of truth.

As anticipated in §1.4, if O is schematically axiomatized there is no result equivalent to Proposition 3.2. Indeed we can actually prove that such a result is impossible. The next two theorems prove this for two well-known systems, Zermelo Fraenkel set theory and Peano Arithmetic. Heck (2009) contains a different proof of the claim in which *PA* is the object theory, although the proof below is somewhat stronger as Heck's theory does not have the full strength of the induction axiom Ind^s.

THEOREM 3.4. $CTD[ZF] \nvdash GRP^s_{ZF}$

We standardly define the cumulative hierarchy V_α and the constructible hierarchy L_α by transfinite induction so that

$$\mathbb{V} = \bigcup_{\alpha \in Ord} V_\alpha \qquad\qquad \mathbb{L} = \bigcup_{\alpha \in Ord} L_\alpha \tag{5}$$

where *Ord* denotes the class of all ordinals. It is well-known that $L_\omega = V_\omega$.

Proof. We will show that $CTD[ZF]$ does not prove the syntactic consistency statement Con^s_{ZF}. The unprovability of GRP^s_{ZF} follows immediately from this fact by modus tollens. Assume for contradiction that $CTD[ZF]$ proves Con^s_{ZF}. By compactness, there is a finite subset of $CTD[ZF]$ in which only finitely many occurrences of the axiom schemas of $CTD[ZF]$ occur—that is of axiom Ind^s, separation and replacement in $\mathscr{L}_\in$. Let us call $CTD^*[U]$ this system, where U is the resulting finite subsystem of *ZF*. By the reflexivity of *ZF*, there is, provably within *ZF*, some α such that $V_\alpha \models U$. From V_α we construct a many-sorted structure $\mathscr{D}$ modeling $CTD^*[U]$, by combining it with a model L_ω for our syntax S, a model for the sequences of variable assignments ${}^\omega V_\alpha$ (i.e. the set of finite sequences of elements of V_α), and by interpreting the function $a(i)$ as the function that applied to a sequence s of elements of V_α yields its i^{th} element.[17] The extension of the satisfaction predicate is then standardly defined as the set

$$\mathscr{E}_D = \{\langle x, y\rangle : y = \sharp\varphi \wedge x \in {}^\omega V_\alpha \wedge V_\alpha \models \varphi[x]\}$$

Now by assumption, we have, provably in *ZF*, that $\mathscr{D} \models Con^s_{ZF}$ and, in particular, that $L_\omega \models Con^s_{ZF}$, as all quantifiers in Con^s_{ZF} are of syntactic sort. Now the relation $V_\omega \models Con^s_{ZF}$ is Δ_1^{ZF},[18] and thus absolute across transitive models. Since we may safely assume that $V_\omega \subset V_\alpha$, we can also conclude that V_α satisfies the relativization to V_ω of the syntactic consistency statement for *ZF*, that is the actual consistency statement for *ZF* in the sense of V_α. Now by the first-order reflection principle for *ZF*, we conclude that $ZF \vdash Con^o_{ZF}$, which yields a contradiction by Gödel's second incompleteness theorem. □

A similar result, although obtained by resorting to a different proof, holds for the case in which the object theory is *PA*. We first introduce an essential component of the argument:

DEFINITION 3.5. *ACA_0 is the system formulated in the language $\mathscr{L}_2$ of second-order arithmetic whose axioms are the basic axioms of PA, the induction axiom*

$$0 \in X \wedge \forall n(n \in X \rightarrow Sn \in X) \rightarrow \forall n(n \in X) \tag{Ind^2}$$

---

[17] Here we are employing the well-known fact that defining an interpretation from the language $\mathscr{L}_T$ of the theory T into the theory S is equivalent to defining in each model $\mathscr{M} \models S$ a model $\mathscr{N} \models T$ with $\mathscr{N} \subset \mathscr{M}$.

[18] Actually $\Delta_1^{KP\omega}$.

where n is a number variable and X is a set variable, and the arithmetical comprehension scheme

$$\exists Y \forall n(n \in Y \leftrightarrow \varphi(n)) \tag{6}$$

where again Y is a set variable and $\varphi(n)$ does not contain second-order quantification nor free occurrences of Y.

THEOREM 3.6. *$CTD[PA] \nvdash GRP^s_{PA}$*

Proof. As before we assume for contradiction that $CTD[PA]$ proves Con^s_{PA}. There is thus a finite susbsystem $CTD^*[U]$ of $CTD[PA]$ which is consistent with Con^s_{PA}. Now the result is a consequence of the following lemma. Let ACA^M_0 be the system whose axioms are exactly the axioms of ACA_0 but the language $\mathscr{L}_2$ of second-order arithmetic is expanded by a set constant M which is allowed to appear into the comprehension scheme of ACA_0. We express with $Mod_X(Y)$ the predicate, definable within ACA_0, that signifies that Y is a model of the set X of sentences of a first-order language.[19]

LEMMA 3.7. *$\mathrm{CT}D^*[U]$ is interpretable in $ACA^M_0 + U + Mod_U(M)$.*

Proof. We first define in $ACA^M_0 + U + Mod_U(M)$ the range of $\mathfrak{L}$ variables. In particular, resorting to the official list of variables of $\mathfrak{L}$—i.e., by denoting with v^o_i variables from $\mathscr{L}_U$, with v^s_i individual variables of $\mathscr{L}_S$, and with v^{sq}_i sequences of $\mathscr{L}_{SQ}$—we will force our translation π to have some built-in mechanism for renaming bound variables such that v^o_i is always renamed as v_{3i}, v^s_i as v_{3i+1} and v^{sq}_i as v_{3i+2}. We then define the range of variables of $\mathfrak{L}$ by means of the formulas

$$v_{3i} \in |M|; \qquad v_{3i+1} = v_{3i+1}; \qquad v_{3i+2} \in {}^{\omega}|M|,$$

where $|M|$ is the domain of the model M of U and ${}^{\omega}|M|$ denotes the set of codes of finite sequences of M-objects, a set definable in ACA_0 by arithmetical comprehension. Clearly $ACA^M_0 + U + Mod_U(M)$ shows the formulas above define nonempty domains. Besides identity, which is preserved in the translation, π assigns to each nonlogical symbol of $\mathfrak{L}$ a corresponding formula in the target theory such that:

(i) to each n-ary function symbol f of $\mathcal{L}_U$ is assigned a formula $\varphi_f(v_3, \ldots, v_{3n}, v_{3(n+1)})$ of $\mathcal{L}^o_2$ such that $ACA^M_0 + U + Mod_U(M)$ proves:

$$\forall v_3, \ldots, v_{3n} \in |M| \exists! v_{3(n+1)} \in |M|(\varphi_f(v_3, \ldots, v_{3n}, v_{3(n+1)}))$$

So for instance to the binary function symbol S^o of $\mathscr{L}_U$ $\pi(\cdot)$ assigns the formula $Sv_3 = v_6$, where v_3, v_6 are new variables ranging over elements of $|M|$ replacing v^o_1, v^o_2.

(ii) similarly, to each n-ary function symbol g of $\mathscr{L}_S$ is assigned a formula of $\mathscr{L}^o_2$ such that

$$ACA^M_0 + U + Mod_U(M) \vdash \forall v_{3+1}, \ldots, v_{3n+1} \exists! v_{3n+4}(\psi_g(v_{3+1}, \ldots, v_{3n+1}, v_{3n+4}))$$

(iii) to the function symbol D of type $\langle sq, s\rangle \to o$ of $\mathfrak{L}$ the formula $val(v_{3i+2}, v_{3i+1}) = v_{3i}$, where *val* designates the arithmetical evaluation function, such that $ACA^M_0 + U + Mod_U(M)$ proves:

$$\forall v_{3i+2} \in {}^{\omega}|M| \forall v_{3i+1} \exists! v_{3i} \in |M|(val(v_{3i+2}, v_{3i+1}) = v_{3i})$$

[19] See Simpson (2009) for the definition of $Mod_X(Y)$ in RCA_0.

(iv) to the predicate *Sat* of $\mathfrak{L}$, $\pi(\cdot)$ assigns the $\mathscr{L}_2^o$-formula $M(v_{3i+1}[v_{3i+2}]) = 1$, where $x[y]$ expresses the result of formally substituting each occurrence of the i^{th} free variable in the formula coded by x by the i^{th} element of the finite sequence of elements encoded by y and

$$M\colon \mathit{Term}_{\mathscr{L}_U^+} \cup \mathit{Sent}_{\mathscr{L}_U^+} \to |M| \cup \{0, 1\}$$

is the function contained in the definition of the model M for the language $\mathscr{L}_U^+$ resulting from $\mathscr{L}_U$ by expanding it with as many constants as the elements of $|M|$ are (cfr. Simpson (2009, §II,8)), with $\mathit{Term}_{\mathscr{L}_U^+}$ and $\mathit{Sent}_{\mathscr{L}_U^+}$ are the set of terms and formulas of $\mathscr{L}_U^+$ as formalized in ACA_0^M.

(v) propositional connectives commute with $\pi(\cdot)$ in the obvious way. For quantification:

$$\pi(\forall v_i^o \varphi) := \forall v_{3i} \in |M| \pi(\varphi)$$
$$\pi(\forall v_i^s \varphi) := \forall v_{3i+1} \pi(\varphi)$$
$$\pi(\forall v_i^{sq} \varphi) := \forall v_{3i+2} \in {}^{\omega}|M| \pi(\varphi)$$

We can thus prove that the translation π actually supports a relative interpretation of $CTD^*[U]$ into $ACA_0^M + U + Mod_U(M)$ by showing that for all sentences $\sigma \in \mathfrak{L}$,

$$CTD^*[U] \vdash \sigma \Rightarrow ACA_0 + U + Mod_U(M) \vdash \pi(\sigma). \tag{7}$$

This is established by induction on the length of a proof in $CTD^*[U]$. It is worth emphasizing that, in the case of the finitely many instances of axiom Ind^s, we have enough arithmetical comprehension to prove the translation of those induction axioms. □

Back to the proof of Theorem 3.6, by instantiating Con_{PA}^s within (7), we get

$$ACA_0^M + U + Mod_U(M) \vdash \pi(Con_{PA}^s) \tag{8}$$
$$ACA_0 + U + \exists Y Mod_U(Y) \vdash \pi(Con_{PA}^s) \tag{9}$$
$$ACA_0 + U \vdash \exists Y Mod_U(Y) \to \pi(Con_{PA}^s) \tag{10}$$

The passage from (8) to (9) is made possible only because the satisfaction predicate, and thus M according to our translation, does not occur in Con_{PA}^s. Now since $U \subseteq PA$, then also $ACA_0 + U \subseteq ACA_0$ and $ACA_0 \vdash Con_U^o$, as U is finite. Therefore, by the arithmetized completeness theorem formalized in ACA_0,[20] we have

$$ACA_0 \vdash \exists Y Mod_U(Y),$$

which combined with (10) yields $ACA_0 \vdash \pi(Con_{PA}^s)$. But $\pi(Con_{PA}^s)$, by looking at the definition of $\pi(\cdot)$ above, is, up to α-conversion, just Con_{PA}^o, which contradicts Gödel's second incompleteness theorem. □

Theorem 3.4 and Theorem 3.6 show how the setting with disentangled syntax actually dissipates the worries expressed in §1.3. Let

$$TAX := \forall a \forall k (Ax_O^s(k) \to Sat(a, k))$$

By inspection of the proof of Proposition 3.2 we may also notice the following:

[20] Cfr. Simpson (2009, §IV. 3).

PROPOSITION 3.8. *Let O be schematically axiomatized, then*

$$CTD[O] + TAX \vdash GRP^{s}_{O}$$

The formalization, within S, of the claim that all axioms of O are true (with O schematically axiomatized), by circumventing the lack of interaction between syntactic and mathematical schemas, allows us to obtain the syntactic global reflection for O.

In the next section, we show further interesting properties of $CTD[O]$.

3.2. Conservativity. We begin the present subsection with a couple of useful notions.

DEFINITION 3.9. *(Expansion of a model). Let $\mathscr{M} = (|\mathscr{M}|, \mathscr{I})$ be a model of the first-order language $\mathscr{L}$. Moreover, let $\mathscr{L}' = \mathscr{L} \cup X$, where X is a set of new nonlogical symbols. We call the resulting $\mathscr{L}'$-structure $\mathscr{M}' = (|\mathscr{M}|, \mathscr{I} \cup \mathscr{I}')$, where $\mathscr{I}'$ is an interpretation of symbols in X, an* expansion *of $\mathscr{M}$.*

DEFINITION 3.10. *(Conservativity). A theory T in $\mathscr{L}_T$ is* proof-theoretically conservative *over the theory S in $\mathscr{L}_S \subseteq \mathscr{L}_T$ if and only if for each theorem φ of T in $\mathscr{L}_S$, already $S \vdash \varphi$. The theory T is* model-theoretically conservative *over the theory S if and only if any model of S can be expanded to a model of T.*

Model-theoretic conservativity implies proof-theoretic conservativity, but the converse does not always hold.[21] $CTD[O]$, regardless the choice of O—it can be either finitely or schematically axiomatized—has the interesting property of being model-theoretically conservative over the object theory O. The recent literature on the philosophy of truth, in fact, contains many attempts to tie up the conservativity of the theory of truth over the underlying object theory to philosophical claims concerning the metaphysical status of the truth predicate. A brief discussion of the philosophical interest of the conservativity of $CTD[O]$ will be contained in the conclusion, although a more thorough treatment will be deferred to a forthcoming work.[22]

THEOREM 3.11. *(Halbach). $CTD[O]$ is model-theoretically conservative over O.*

Proof. The key detail of the proof, which also makes the disentangled setting different from the usual one, is represented by the possibility of employing a standard interpretation of the syntax and of the length of the sequences of variable assignments of *SQ by stipulation*, even in the presence of a non-standard model of O. We thus take $\mathbb{N}$ to model S. Now we construct the model $\mathscr{D}$ of $CTD[O]$. Syntactic vocabulary is interpreted standardly on $\mathbb{N}$. Sequences are interpreted as ranging over the set ${}^{\omega}|\mathscr{M}|$. The function $.(.)$—and consequently D, if function symbols are present in $\mathscr{L}_O$—is naturally interpreted as a function

$$f^{\mathscr{D}} : {}^{\omega}|\mathscr{M}| \times \mathbb{N} \to |\mathscr{M}|$$

---

[21] For instance, the theory *UTB*, formulated in $\mathscr{L} \cup \{T\}$ by adding to the axiom of *PA* the schema, for any $\varphi(v)$

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

is proof-theoretic conservative over *PA* but not model-theoretically conservative over it. A simpler example concerns $Th(\mathbb{N})$ in $\mathscr{L}$ and $S = Th(\mathbb{N}) \cup \{c > \overline{n} \mid n \in \omega\}$. Clearly not every model of $Th(\mathbb{N})$ can be expanded to a model of S, but yet S is a conservative extension of $Th(\mathbb{N})$ as in a proof of $\varphi \in \mathscr{L}$ from S the constant c can be replaced by a term $\bar{m}$ determined by the finitely many axioms $c > \bar{n}$ occurring in the proof.

[22] See also Nicolai (forthcoming).

from the cartesian product of the set of finite sequences of $\mathscr{M}$-objects and of the standard model of arithmetic into the domain of $\mathscr{M}$. Finally, the extension of the satisfaction predicate is defined standardly as:

$$\mathscr{E}_D = \{\langle x, y\rangle : y = \sharp\varphi \in \mathbb{N} \text{ for some } \varphi \text{ from } \mathscr{L}_O, x \in {}^{\omega}|\mathscr{M}| \text{ and } \mathscr{M} \vDash \varphi[x]\} \qquad (11)$$

where again $\sharp\varphi$ denotes the Gödel number of the formula φ. A complete definition of $\mathscr{E}_D$ would of course be carried out on the complexity of φ, but this creates no problems as syntactic objects are of standard length. We have thus constructed the expansion $(\mathscr{D}, \mathscr{E}_D)$ of $\mathscr{M}$. We see that $(\mathscr{D}, \mathscr{E}_D) \vDash CTD[O]$. Axioms of O and of S are clearly satisfied by $\mathscr{D}$. SQ and Ind^s satisfied by $\mathscr{D}$ by a straightforward inductive argument in the metatheory on the (standard) length of sequences and syntactic objects respectively. Axioms for satisfaction follow from the definition of $\mathscr{E}_D$. □

COROLLARY 3.12. *$CTD[O]$ is proof-theoretically conservative over O.*

COROLLARY 3.13. *$CTD[O] + TAX$ is model-theoretically, and thus proof-theoretically conservative over the schematically axiomatized theory O.*

3.3. Extending schemas of O. When the object theory O is schematically axiomatized, $CTD[O]$ is constructed in such a way that schemas of O are only allowed to contain formulas of $\mathscr{L}_O$. In Section 1 we already motivated this restriction by showing how the interaction between *syntactic* and *mathematical* instances of the schemas of O is the main responsible for the nonconservativity of the truth theory over O. However, many authors would find this restriction suspicious: once one has accepted schemas in the object language, she is also committed to accepting instances of those schemas in each expansion of the base language by new (interpreted) predicates.[23] In the present subsection we investigate the possibility of allowing arbitrary formulas belonging to the vocabulary of $\mathfrak{L}$ to occur into schemas of O. This move would make the resulting theories nonconservative over O.

Let us consider the schemas

$$\varphi(0^o) \wedge \forall x(\varphi(x) \to \varphi(S^o x)) \to \forall x(\varphi(x)) \quad \text{for all } \varphi \in \mathfrak{L} \qquad (Ind^+)$$

$$\forall x \exists y \forall z(z \in y \leftrightarrow z \in x \wedge \varphi(z)) \quad \text{for any } \varphi \in \mathfrak{L} \qquad (Sep^+)$$

$$\forall x \exists! y \varphi(x, y) \to \forall x \exists y \forall u\big(u \in y \leftrightarrow (\exists z \in x)\varphi(z, y)\big) \quad \text{for any } \varphi \in \mathfrak{L} \qquad (Rep^+)$$

In the present subsection we will look at the following theories:

$$CTD[PA]^+ := CTD[PA] + Ind^+$$

$$CTD[ZF]^+ := CTD[ZF] + Sep^+ + Rep^+$$

THEOREM 3.14. *Let O be either PA or ZF. Then $CTD[O]^+$ is not a conservative extension of O.*

Proof. We give the argument for $CTD[PA]^+$. The strategy for $CTD[ZF]^+$ is specular.

Working within *PA*, we start with a suitable partial truth predicate $Tr^o_{\Sigma_n}(x)$ which applies to codes of sentences of the language $\mathscr{L}$ of *PA* represented in $\mathscr{L}$ itself. It is important

[23] Cfr. for instance McGee (1997) and Lavine (1999). In Nicolai (forthcoming) it is defended the position that, even if an *expansionist* stance is taken towards schemas of a theory O, one is not committed to the expansions of those schemas to the entire vocabulary of $CTD[O]$.

to recall, at this stage, our indexing of syntactic objects. Since we are now formalizing syntactic operations and notions on and about $\mathscr{L}$ within *PA* itself, we index them with the superscript o and we place a line over codes of $\mathscr{L}$-expressions instead of putting them into Gödel corners. It is well-known that, for $\varphi(x)$ a Σ_n-formula of $\mathscr{L}$,

$$Tr^o_{\Sigma_n}(\overline{\varphi(\dot{x})}) \leftrightarrow \varphi(x) \tag{12}$$

where $\varphi(\dot{x})$ is short for $sub^o(\overline{\varphi}, \overline{x}, num^o(x))$, that is the result of formally substituting the free variable x in the formula φ with the numeral for x. We claim that

$$CTD[PA]^+ \vdash \forall x\big(Sent^o_{\mathscr{L}}(x) \wedge Bew^o_{PA}(x) \rightarrow \forall a \forall k(a(k) = x \rightarrow \exists n(Sat(a, \ulcorner Tr_{\Sigma_n}(v_k)\urcorner)))\big). \tag{13}$$

(13) is provable by induction on the length of the proof in *PA*, with a subsidiary induction to establish the base case, both of which are available in $CTD[PA]^+$. However, we should make sure that the following holds for all $\varphi(v)$ in $\mathscr{L}$:

$$\forall a \forall \ulcorner s \urcorner \forall \ulcorner t \urcorner \big(Sent^s_{\mathscr{L}}(\dot{\forall} v \varphi) \wedge D(a, \ulcorner s \urcorner) = D(a, \ulcorner t \urcorner) \rightarrow (Sat(a, \varphi[s/v]) \leftrightarrow Sat(a, \varphi[t/v]))\big) \tag{14}$$

(14) is proved by applying Ind^s to the formula

$$\forall \ulcorner \varphi(v_i) \urcorner \forall n(Sent^s_{\mathscr{L}}(\dot{\forall} v_i \varphi) \wedge lc^s(\ulcorner \varphi(v_i) \urcorner) \leq n \rightarrow \forall a \forall \ulcorner s \urcorner \forall \ulcorner t \urcorner \big(D(a, \ulcorner s \urcorner) = D(a, \ulcorner t \urcorner) \rightarrow (Sat(a, \varphi[s/v]) \leftrightarrow Sat(a, \varphi[t/v]))\big)\big) \tag{15}$$

where $lc^s(\cdot)$ represents in $\mathscr{L}_S$ the primitive recursive function that takes a formula of $\mathscr{L}$ and returns the number of its logical symbols. In the proof of (13), there are two non trivial cases in which (14) plays an important role. The first of these concerns the logical axioms of *PA*. The second concerns the induction axioms of *PA*. Given an arbitrary instance of induction, x, we require to show that

$$\forall a \forall k(a(k) = x \rightarrow \exists n(Sat(a, \ulcorner Tr_{\Sigma_n}(v_k)\urcorner))). \tag{16}$$

This follows, however, from the application of Ind^+ to the formula:

$$\psi(y) := \forall i, j\ \forall a \exists n(Fml^o(a(i)) \wedge a(j) = y \wedge Sat(a, \ulcorner Tr_{\Sigma_n}(sub^o(v_i, \overline{v}, v_j))\urcorner)). \tag{17}$$

By instantiating an absurdity ξ of O within (13), we obtain $\neg Bew^o_{PA}(\overline{\xi})$, i.e. Con^o_{PA}. □

The way in which $CTD[O]^+$ is constructed is somewhat unnatural, at least from our point of view, as the interaction between 'mathematical' and 'syntactic' schemas, essential to prove (13), was exactly what the setting with 'disentangled syntax' wanted to avoid. At any rate, Theorem 3.14 stresses once more how essential the identification between syntax and object theory is in order for our theory of truth to be stronger than its schematically axiomatized object theory.

3.4. Coding axioms. A further strengthening of $CTD[O]$ concerns the addition of suitable axioms extending our bridge laws connecting the domain of the object theory O and the realm of syntactic objects. As we shall argue in the concluding section, this enrichment of $CTD[O]$ amounts to a faithful formalization of our informal metamathematical practice. The strategy, suggested in Halbach (2011), is reminiscent of the procedure employed in the translation of the formalization of a two-sorted scientific theory containing variables for mathematical entities and for physical entities into its nominalistically accept-

able version, as presented in Burgess & Rosen (1997).[24] Our aim is to extend $CTD[O]$, when O is finitely axiomatized, and $CTD[O]+TAX$, when O is schematically axiomatized, in such a way that any sentence of the language of S not containing semantic vocabulary which is provable from $CTD[O]$ and $CTD[O] + TAX$ gets mapped into an equivalent statement in the language of O. In particular, if a sentence A of $\mathscr{L}_S$ contains syntactic information regarding $\mathscr{L}_O$, we want A to be mapped into a sentence σ of $\mathscr{L}_O$ conveying the same syntactic content concerning $\mathscr{L}_O$. In particular, if our theory of truth with disentangled syntax proves Con^s_O, we want our coding axioms to be able to define a translation τ from the syntactic language $\mathscr{L}_S$ into the object-language $\mathscr{L}_O$ such that $\tau(Con^s_O)$ turns out to be provably equivalent—within our extensions of $CTD[O]$ and $CTD[O] + TAX$—to Con^o_O, i.e. the consistency statement for O as formalized in $\mathscr{L}_O$ itself.

In order to obtain the desired strengthenings of $CTD[O]$ and $CTD[O] + TAX$, it suffices to start from their subtheory $S[O]$.

DEFINITION 3.15. *$S[O]$ is the two-sorted subtheory of $CTD[O]$ featuring quantification over the syntactic and the mathematical domain and whose axioms are the basic axioms of S, the axioms of O and the syntactic induction $Ind^s\restriction$ where:*

$$Ind^s\restriction := Ind^s \text{ restricted to the language of } \mathscr{L}_{S[O]}.$$

The formulation of the axioms of the theory $S[O]^{ca}$, which is obtained from $S[O]$ by adding suitable coding axioms to it, is sensitive to the choice of the language of O. In particular, in order to obtain $S[O]^{ca}$, we will expand the language $\mathscr{L}_{S[O]}$ with a binary predicate symbol C of type $\langle o, s\rangle$ allowed to appear into $Ind^s\restriction$ and add to $S[O]$ one of the two sets of axioms presented in Tables 2 and 3.

Axiom $CA3$, in both cases, is particularly required as the intended model of the syntax S is always $\mathbb{N}$: this means that—without $CA3$—in the presence of a nonstandard model of our object theory, we would not have enough resources, in our disentangled syntax,

Table 2. *Coding axioms for $\mathscr{L}_O$ is $\mathscr{L}$*

$CA1$	$\forall x(C(x, 0^s) \leftrightarrow x = 0^o)$
$CA2$	$\forall x \forall n\big(C(x, n) \rightarrow C(S^o x, S^s n)\big)$
$CA3$	$\forall x \exists! n C(x, n)$

Table 3. *Coding axioms for $\mathscr{L}_O$ is $\mathscr{L}_\in$*

$CA1$	$\forall x(C(x, 0^s) \leftrightarrow x = \varnothing)$
$CA2$	$\forall x \forall n\big(C(x, n) \rightarrow C(x \cup \{x\}, S^s n)\big)$
$CA3$	$\forall x \exists! n C(x, n)$
$CA4$	$\forall x \forall n(C(x, n) \rightarrow x \in \omega)$

to talk about all the objects in the domain of O. In particular, formulas obtained by quantification within $\mathscr{L}_O$ over 'all sentences of $\mathscr{L}_O$', 'all proofs of O' or, more in general, syntactic notions which are at least numeralwise represented in S would suffer a fatal shift

[24] Thanks to Jeffrey Ketland for giving us access to an unpublished note suggesting this similarity.

in meaning when translated in their equivalent in $\mathscr{L}_S$.[25] However, in the presence of the other coding axioms, as we will see shortly, *CA3* forces the interpretation of the syntax formalized within O obtained from the translation of the 'disentangled' syntax via coding axioms to be standard.

Under a suitable choice of O, $S[O]^{ca}$ would prove the equivalence of every syntactic claim with an object-theoretic equivalent. To obtain this, we first notice that $C(x, n)$ so defined is a *representation* formula for $S[O]$ in the sense of Burgess & Rosen (1997), namely it has, besides the property expressed by *CA*3, also the property

$$\forall n \exists! x \; C(x, n) \tag{18}$$

This is established in $S[O]^{ca}$ by an instance of $Ind^s\restriction$. The next step is to definitionally extend $S[O]^{ca}$ so to have, for each atomic formula of $\mathscr{L}_S$, a so-called *counterpart formula*:[26]

$$\psi_{0^s}(x) := \exists k(C(x, k) \wedge k = 0^s) \tag{$CF0^s$}$$
$$\psi_{S^s}(x, y) := \exists k \exists l(C(x, k) \wedge C(y, l) \wedge S^s k = l) \tag{CFS^s}$$
$$\psi_{+^s}(x, y, z) := \exists k \exists l \exists m(C(x, k) \wedge C(y, l) \wedge C(z, m) \wedge k +^s l = m) \tag{$CF+^s$}$$
$$\psi_{\times^s}(x, y, z) := \exists k \exists l \exists m(C(x, k) \wedge C(y, l) \wedge C(z, m) \wedge k \times^s l = m) \tag{$CF\times^s$}$$

THEOREM 3.16. *Let O be a theory in the language of arithmetic or set theory. Then for every formula $A(k_1, \ldots, k_n)$ of $\mathscr{L}_S$ with exactly n free variables, there is a $\mathscr{L}_O$-formula $A^*(x_1, \ldots, x_n)$ with the same number of free variables such that*

$$S[O]^{ca} \vdash \forall k_1, \ldots, k_n \forall x_1, \ldots, x_n \big(C(x_1, k_1) \wedge \cdots \wedge C(x_n, k_n) \rightarrow (A(k_1, \ldots, k_n) \leftrightarrow A^*(x_1, \ldots, x_n))\big).$$

Proof. The proof undergoes some minor changes if O is formulated in the language of arithmetic or if it is formulated in the language of set theory. We consider the first, more tedious case as function symbols are present in $\mathscr{L}_O$, but it should be immediate to see how to proceed in the simpler case.

We define a translation τ of $\mathscr{L}_S$ into $\mathscr{L}_O$ as follows:

(i) variables occurring in $\mathscr{L}_S$-formulas under the translation τ are taken to range over the domain of O. As in the proof of Lemma 3.7, we can assume that our interpretation maps each variable v_i^s of $\mathscr{L}_S$ to the $\mathscr{L}_O$-variable v_{3i+1} or allow ourselves with some mechanism for α-conversion.

(ii) τ preserves identity.

(iii) To the symbols 0^s, S^s, $+^s$ and $\times^s$ of $\mathscr{L}_S$, τ assigns the $\mathscr{L}_O$-symbols 0^o, S^o, $+^o$ and $\times^o$.

(iv) Finally, τ commutes with $\neg$, $\wedge$ and

$$\tau(\forall k A) := \forall x(\tau(A))$$

[25] That we would always have sentences, proofs of nonstandard length in a nonstandard model of O is an immediate consequence of the Overspill Principle.

[26] Cfr. Burgess & Rosen (1997, p.87).

The translation τ forms the basis for defining the formula A^*. We do this by proving

$$\forall x(\psi_{0^s}(x) \leftrightarrow x = 0^o) \tag{19}$$

$$\forall x, y(\psi_{S^s}(x, y) \leftrightarrow S^o x = y) \tag{20}$$

$$\forall x, y, z(\psi_{+^s}(x, y, z) \leftrightarrow x +^o y = z) \tag{21}$$

$$\forall x, y, z(\psi_{\times^s}(x, y, z) \leftrightarrow x \times^o y = z) \tag{22}$$

Now (19) follows immediately from *CA*1. For (20)—right-to-left direction—we start from $S^o x = y$. By *CA*3 there is a unique l such that $C(y, l)$ and thus also $C(S^o x, l)$. But l must have the form of $S^s k$ for some k as otherwise l is 0^s, and thus, by *CA*1, $S^o x = 0^o$, which is impossible, thus $S^s k = l$. On the other hand, from $\psi_{S^s}(x, y)$ by *CA*2 we have $C(S^o x, S^s k)$ for some k and $C(y, S^s k)$, whence $S^o x = y$ by (18).

For (21), we first prove by Ind^s on l the universal closure of

$$C(x, k) \wedge C(y, l) \rightarrow C(x +^o y, k +^s l) \tag{23}$$

The base case is a consequence of (19). For the induction step, we argue as follows: suppose $l = S^s l_0$, by (18) $C(y_0, l_0)$ holds for some y_0. By (20) we deduce that $S^o y_0 = y$. But now we can apply the induction hypothesis $C(x +^o y_0, k +^s l_0)$. Now by *CA*2 we obtain $C(S^o(x +^o y_0), S^s(k +^s l_0))$, as required. From (23) we get (21) by applying *CA*3 and (18).

The remaining case is similar to the case of addition. By means of (19)–(22) we see that for each formula $A(k_1, \ldots, k_n)$ of the syntax theory S there is a formula $A^*(x_1, \ldots, x_n)$ of $\mathscr{L}_O$ such that

$$\forall k_1, \ldots, k_n \forall x_1, \ldots, x_n \big(C(x_1, k_1) \wedge \cdots \wedge C(x_n, k_n) \rightarrow (A(k_1, \ldots, k_n) \leftrightarrow A^*(x_1, \ldots, x_n))\big) \tag{24}$$

(24) is established by induction on the complexity of the $\mathscr{L}_S$-formula $A(k_1, \ldots, k_n)$. The atomic case is covered by (19)–(22); the cases of the propositional connectives are straightforward as well, given the definition of τ; for the universal quantifier, i.e. when $A(k_1, \ldots, k_n)$ is $\forall k_j B(k_j, k_1, \ldots, k_n)$, with $j > n$, we have $B^*(x_j, x_1, \ldots, x_n)$ by induction hypothesis, and by (18) we can safely generalize over the x_j, as there is no other occurrence of x_j in $B^*(x_1, \ldots, x_n)$. □

COROLLARY 3.17. *For every sentence $A \in \mathscr{L}_S$:*

$$S[O]^{ca} \vdash A \leftrightarrow \tau(A) \tag{25}$$

Let us take into consideration the theory $CTD[O]^{ca}$, which extends $S[O]^{ca}$ by expanding its language, as we know, with a new sort of variables for sequences of variable assignments, with *Sat*, .(.) and D, and by extending its set of axioms in line with what is displayed in Table 1. Given Theorem 5 and Corollary 3.17, we have:

PROPOSITION 3.18. *If O is a finitely axiomatized theory then $CTD[O]^{ca} \vdash Con^o_O$. If O is schematically axiomatized then $CTD[O]^{ca} + TAX \vdash Con^o_O$.*

In the next section we try to give a first philosophical assessment of the results obtained.

§4. Concluding Remarks. We tried to explain why the usual way of constructing formal theories of truth, in which the theory formalizing the syntax of the language of an object theory O is contained in O itself is at best the crystallization of a widespread habit among logicians, but it cannot be counted as the unique way in which our theories

can be constructed. On the contrary, we tried to provide arguments belonging to different categories in support of the thesis that a proper axiomatization of the notion of truth for O may also be set up to contain, á la Tarski, a distinct class of 'linguistic' entities denoting structural-descriptive relations among expressions of $\mathscr{L}_O$. In Section 2 we presented the theory $CTD[O]$, and in Section 3 we measured its strength and examined the properties of the theories $CTD[O]^+$ and $CTD[O]^{ca}$ which are obtained from $CTD[O]$ by adding some extra-resources to it. The very formulation of $CTD[O]$ and the results provable from its axioms seem to be relevant for philosophical purposes at least in two senses: firstly, the results on $CTD[O]$ can contribute to the debate about the explanatory power of the notion of truth; specifically in the context of the discussion of the so-called 'conservativeness argument' against deflationism (§4.1); moreover, $CTD[O]^{ca}$ offers a realization of our informal metamathematical discussion as described in §1.1 and Halbach (2011). This latter point will be discussed in §4.2.

4.1. Conservativeness and Syntax. We refer to the so-called conservativeness argument as the challenge to the deflationary conception of truth formulated in (Horsten 1995), (Shapiro 1998) and (Ketland 1999). The argument stems from the provability in *CT* of the Global Reflection Principle for *PA* and consequently of Con^o_{PA}.[27] Deflationism holds that the notion of truth is metaphysically weak, in the sense that it is not a property in a genuine sense and it does not play a substantial role in philosophical and scientific explanations. According to the conservativeness argument, if by only adding axioms characterizing the notion of truth to *PA* we were able to prove new theorems in the language of *PA*, then the notion of truth could not be counted as 'insubstantial'. As Shapiro puts it:

> How thin can the notion of arithmetic truth be, if by invoking it we can learn more about natural numbers? By taking on this supposedly insubstantial, mere device for indirect endorsement, we can establish facts about the natural numbers we could not establish before. (Shapiro, 1998, pp. 499–500)

From the delationist's side, Field has argued that the reason why we 'learn more about natural numbers' resides in the fact that, by extending the induction axioms of *PA* to contain semantic vocabulary, we are actually enhancing its mathematical potential, and there is no surprise if, by combining the resulting theory with compositional axioms for truth, we obtain a conservative extension of *PA*. Moreover, the theory $CT\!\upharpoonright$, which is *CT* with induction restricted to arithmetical formulas, is still conservative over *PA*. Therefore, there is a sense in which the deflationist can accept the conservativeness requirement for deflationary acceptable theories without giving up her doctrine.

Although Field's idea has raised many objections,[28] it captures an essential point already outlined in §1.3: as remarked in Heck (2009), in the proof of the global reflection principle for *PA* in *CT*, there are two ways in which the induction of *PA* is employed, one *syntactic* and one *mathematical*, as exemplified by the two instances (1) and (2). The only use of induction that we want to allow in our theory of truth is the syntactic one, as the mathematical one rests upon the fact that the syntax is contained in the object

---

[27] Notice that, following the notation already employed in the paper, we write Con^o_{PA}, as it is a consistency statement formalized in the language of *PA* itself.

[28] For instance, as remarked in Horsten (2011), if we consider the extended induction axioms of *CT* as arithmetical axioms, then we might also consider, as base theory of *CT*, the theory *PAT*, that is *PA* formulated in the language $\mathscr{L} \cup \{T\}$. If we go this way, however, *CT* would still be a non conservative extension of *PA*.

theory and, for reasons that should be clear at this stage, we do not want to depend on this simplification in providing a sufficiently general framework for truth, at least for the specific purpose of measuring the strength of principles of truth as opposed to syntactic and mathematical principles. Moreover, some of the results presented in the previous sections seem to support Field's tenet: in particular, Proposition 3.2 and Proposition 3.8 tell us that the 'syntactic' global reflection principle for O—GRP^s_O—respectively when O is finitely or schematically axiomatized, is provable in our theories of truth with disentangled syntax $CTD[O]$ and $CTD[O] + TAX$. However, both theories are still conservative over O. This should count as a point in favor of Field's idea, although in a very distinctive sense, namely, if we formulate things properly, without adding new mathematical power to our object theory O, our theory of truth will always be a conservative extension of O.

We will now just hint to the reasons why we think that (i) the results of this paper do not represent an easy way out for the deflationist; and (ii) the original formulation of the conservativeness argument rests upon the assumption that expressions of the object language are also fully-fledged mathematical objects, and that if our way of conceiving the axiomatization of the truth predicate has some advantages with respect to the one that we want to revise, also the conservativeness argument has to be reformulated.

Ad (i): $CTD[O]$ is conservative over O but it is not conservative over the two-sorted theory $S[O]$. Con^s_{PA} is in fact a statement in the language of $S[O]$ which is provable in $CTD[O]$ (or $CTD[O] + TAX$) but not in $S[O]$. It is imaginable to consider now the theory $S[O]$ as the theory over which the deductive power resulting from the addition of the truth predicate has to be evaluated, as truth is added to our disentangled syntax plus our mathematical object theory in the same way as truth was added to the object theory *PA* when the syntax for $\mathscr{L}$ was developed within *PA* itself. Already at the early stage of the debate concerning the conservativeness argument the conservativeness over the underlying ontology of expressions has sometimes been considered as the main target.[29] If the anti-deflationist takes this route, however, she has to make sense of Theorems 3.4 and 3.6 which state that without interactions between the syntactic and the mathematical—permitted in the usual setting by the duplex role of schemata of *PA* and *ZF*—we cannot reproduce in our theory of truth the reflective reasoning supporting our acceptance of the object theory, starting with the essential reflection on the truth of the axioms of O. She might consider the possibility of adding the statement $\forall k \forall a(Ax^s_O(k) \rightarrow Sat(a, k))$ as a new axiom. She might then argue that this is a harmless assumption given that the provability of the Global Reflection Principle, in the usual as well as the disentangled setting, relies on the assumption of the soundness of O, and that schematically axiomatized theories are somehow a special case, whereas in the case of finitely axiomatized theories only Tarski biconditionals, which are provable in $CTD[O]$, are needed to reconstruct our belief in the truth of the axioms of O.

There seems thus to be a way for the anti-deflationist to interpret the results on theories of truth with disentangled syntax in a favorable way. However, a specific effort needs to be made to justify the assumption of the syntactic claim postulating the truth of the axioms of the schematically axiomatized object theory.

Ad (ii): *CT* proves the $\mathscr{L}_{PA}$-sentence Con^o_{PA}. So it proves that a particularly satisfactory axiomatization of our reasoning about natural numbers does not lead to contradiction when combined with classical logic.[30] How can we then "learn more about the natural

[29] Cfr. Halbach (2001) and Shapiro (2002).

[30] Cfr. Hofweber (2000).

numbers" by proving the coded claim Con^o_{PA}? It seems that what we are actually doing is to fully express our commitment in the soundness of this particular axiomatization of number theory known as *PA*, by means of the global reflection principle—and thus by means of our theory of truth—and the consistency statement for *PA* is a consequence of this commitment. But GRP^s_{PA} appears to express exactly this commitment without displaying the peculiarities related to the identification of syntactic operations and functions over expressions of $\mathscr{L}_O$ with numbers or sets in the sense of our mathematical theory O. In particular, we do not need any interaction between the syntactic and the mathematical use of the axiom schema of induction of *PA*—of the kinds shown in Proposition 1.2 in §1.4—to prove GRP^s_{PA}; at the same time, more importantly, $CTD[O]$ or $CTD[O] + TAX$ give us *new syntactic consequences* with respect to $S[O]$. By paraphrasing Shapiro's formulation of the conservativeness argument,[31] before moving to $CTD[PA]+TAX$, it was still logically possible for the axioms of $S[O]$ to be true and yet Con^s_{PA} to be false, but it is not logically possible for the axioms of $CTD[PA] + TAX$ to be true and Con^s_{PA} false. In other words, GRP^s_{PA} has also the advantage of characterizing what our commitment to the acceptance of *PA* is, namely, the formal statement—not necessarily belonging to $\mathscr{L}_{PA}$—declaring the soundness of a particular axiomatization, and of measuring what the exact explanatory power of these principles is, that is a characterization of our reasoning about some portion of mathematics involving truth-theoretic and syntactic principles.

The truth predicate of $CTD[O]$ thus allowed us to 'make commitments (about matters not involving truth) that we could not make without it',[32] without increasing our knowledge of the mathematical structure captured by the theory O. The construction of theories of truth in the style of $CTD[O]$ certainly does not offer a definitive answer to the debate about the alleged 'insubstantiality' of truth. Nevertheless any formulation, or reformulation, of the conservativeness argument should take into account a setting that, by isolating the various components of our metamathematical reflection, seems to support an understanding of the truth predicate as a very special predicate with a pronounced logico-linguistic character.

4.2. Capturing informal metamathematical reasoning. If we establish that a sentence σ of $\mathscr{L}_O$ is provable in the mathematical theory O, in our informal metamathematical discussion we can conceive this claim as a syntactic relation occurring between the axioms of O, its rules of inference and the statement σ. To this relation, given standard metamathematical techniques, correspond the formal statement $Bew^o_O(\overline{\sigma})$. However, it is also usually required that

$$\sigma \text{ is provable in } O \quad \text{if and only if} \quad O \vdash Bew^o_O(\overline{\sigma}) \tag{26}$$

This correspondence is faithfully captured by Corollary 3.17. In $CTD[O]$ the informal claim on the left is formalized in S as $Bew^s_O(\ulcorner \sigma \urcorner)$.

A similar remark holds for peculiar sentences as the Gödel sentence γ for O and in particular for the 'mathematical' consistency statement Con^o_O. In $\mathfrak{L}$ we can express formal versions of the informal claims concerning the soundness of O, embodied by GRP^s_O, and the consistency of O, expressed by the $\mathscr{L}_S$-sentence Con^s_O. Since the latter is a purely syntactic statement, i.e. not containing semantic vocabulary of $\mathfrak{L}$, $CTD[O]^{ca}$ and $CTD[O]^{ca} + TAX$ prove Con^o_O. This is the content of Proposition 3.18.

---

[31] Shapiro (1998, p. 497).

[32] Cfr. Field (1999, p. 534).

We notice that the conservativity of $CTD[O]$ and $CTD[O] + TAX$ over O entails that, without coding axioms, the theories of truth with disentangled syntax cannot capture our common metatheoretic practice as depicted above. On the contrary, $CTD[O]^{ca}$ seems to offer a satisfactory picture of our informal metatheoretic discussion as characterized in Halbach (2011).

§5. Acknowledgments. The research of both authors was supported by the Arts and Humanities Research Council UK AH/H039791/1. The second author also benefited from the support of the CUC scholarship from the Progetto Culturale, CEI.

BIBLIOGRAPHY

Barwise, J. (1975). *Admissible Sets and Structures*. Berlin: Springer

Burgess, J., & Rosen, G. (1997). *A Subject With No Object. Strategies for a Nominalistic Interpretation of Mathematics*. Oxford: Clarendon Press.

Craig, W., & Vaught, W. (1958). Finite axiomatisability using additional predicates. *The Journal of Symbolic Logic*, **23**, 289–308.

Enderton, H. (2001). *A Mathematical Introduction to Logic* (second edition). New York: Harcourt/Academic Press.

Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, **49**, 35–91.

Feferman, S. (1982). Inductively presented systems and the formalization of metamathematics. In Van Dalen, D., Lascar, D., and Smiley, T. J., editors. *Logic Colloquium '80*. Amsterdam: North-Holland, pp. 95–128.

Feferman, S. (1989). Finitary inductively presented logics. In Ferro, R., Bonotto, C., Valentini, S., and Zanardo, A., editors. *Logic Colloquium 1988*. Amsterdam: North-Holland, pp. 191–220.

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

Field, H. (1999). Deflating the conservativeness argument. *Journal of Philosophy*, **96**(10), 533–540.

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

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic* **164** (11), 1484–1523.

Hájek, P., & Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Berlin: Springer.

Halbach , V. (2001). How innocent is deflationism? *Synthese,* **126**(1/2), 167–194.

Halbach, V. (2009). Axiomatic Theories of Truth. In Edward N. Zalta, editor. *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*. Available from: http://plato.stanford.edu/archives/win2009/entries/truth-axiomatic/.

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

Heck, R. (2009). *The Strength of Truth Theories*. Unpublished manuscript. Available from: http://rgheck.frege.org/pdf/unpublished/TruthTheories.pdf.

Hofweber, T. (2000). Proof-theoretic reduction as a philosopher's tool. *Erkenntnis*, **53**, 127–146.

Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In Cortois, P., editor, *The Many Problems of Realism (Studies in the General Philosophy of Science: Volume 3)*. Tilburg: Tilburg University Press, pp. 173–187.

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

Lavine, S. (1999). *Skolem was wrong*. Unpublished manuscript.

Mancosu, P. (1991). Generalizing classical and effective model theory in theories of operations and classes. *Annals of Pure and Applied Logic* **52**, 249–308.

Manzano, M. (1996). *Extensions of First-Order Logic*. Cambridge: Cambridge University Press.

McGee, V. (1997). How we learn mathematical language *The Philosophical Review* **106**, (1), 35–68.

Nicolai, C. (forthcoming). *Axiomatic Truth, Syntax and Deflationism*. Dphil Thesis, Oxford.

Niebergall, K. G. (2011). Mereology. In Horsten, L. & Pettigrew, R., editors. *Continuum Companion to Philosophical Logic*. London: Continuum.

Quine, W. V. (1946). Concatenation as a basis for arithmetic. *Journal of Symbolic Logic,* **11**(4), 105–114.

Shapiro, S. (1998). Truth and proof: through thick and thin. *The Journal of Philosophy,* **95**, 493–521.

Shapiro, S. (2002). Deflation and conservation. In Halbach V. & Horsten L. *Principles of Truth*. Frankfurt: Dr. Hänsel-Hohenhausen.

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

Smoryński, C. (1977). The incompleteness theorems. In Barwise, J., editor. *Handbook of Mathematical Logic.* Amsterdam: North-Holland.

Tarski, A. (1936). The concept of truth in formalized languages. In Woodger, H. J. (translator and editor), *Logic, Semantic, Metamathematics:Papers of Alfred Tarski From 1922-1938*. Oxford: Clarendon Press, 1956, pp. 152–278.

Tarski, A. (1944). The semantic conception of truth. *Philosophical and Phenomenological Research* **4**, 341–375.

Troelstra, A. S., & Schwichtenberg, H. (2000). *Basic Proof Theory*. Cambridge tracts in theoretical computer science, no. 43 (second edition). Cambridge: Cambridge University Press.

Visser, A. (2009). Growing commas. A Study of sequentiality and concatenation, *Notre Dame Journal of Formal Logic,* **50**(1), 61–85.

FACULTY OF PHILOSOPHY
OXFORD, OX2 6GG, UK
E-mail: graham.leigh@philosophy.ox.ac.uk

FACULTY OF PHILOSOPHY
OXFORD, OX2 6GG, UK
E-mail: carlo.nicolai@some.ox.uk